

Thesis submitted in fulfilment of the requirements for the degree

Dr. rer. pol.

on the topic

**Estimation of Disaggregated Indicators with
Application to the Household Finance and
Consumption Survey**



to the

Chair of Applied Statistics

School of Business and Economics

Freie Universität Berlin

submitted by

Ann-Kristin Kreutzmann

born in Meppen

Berlin, 2018

Ann-Kristin Kreutzmann, *Estimation of Disaggregated Indicators with Application to the Household Finance and Consumption Survey*,
October 2018

Supervisors:

Prof. Dr. Timo Schmid (Freie Universität Berlin)

Prof. Nicola Salvati, Ph.D. (Università di Pisa)

Location:

Berlin

Date of defense:

December 19, 2018

Acknowledgements

I would like to express my deep gratitude to my supervisor, Prof. Dr. Timo Schmid (Freie Universität, Germany). His guidance and encouragement have been invaluable for the success of this project.

I am also very thankful to Prof. Nicola Salvati, PhD (Università di Pisa, Italy), for useful and interesting discussions on the topic of this thesis and for his profound statistical input.

Special thanks go to the Deutsche Forschungsgemeinschaft (DFG) for supporting this thesis.

Furthermore, I am very grateful to all others who have accompanied me on the way to this thesis, especially my colleagues at the Chair of Statistics and the Statistical Consulting Unit *fu:stat*.

Needless to say, I am always thankful to my beloved family.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "Estimation of Disaggregated Indicators with Application to the Household Finance and Consumption Survey."

1. Kreutzmann, A.-K., Marek, P., Salvati, N., and Schmid, T. (2018). **The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany**, submitted to the Annals of Applied Statistics.
2. Halbmeier, C., Kreutzmann, A.-K., Schmid, T., and Schröder, C. (2018). **The fayherriot command for estimating small-area indicators**, submitted to The Stata Journal.
3. Kreutzmann, A.-K., (2018). **Estimation of sample quantiles: Challenges and issues in the context of income and wealth distributions**, submitted to AStA Wirtschafts- und Sozialstatistisches Archiv, major revision. The final article is published in AStA Wirtschafts- und Sozialstatistisches Archiv, Volume 12, Issue 3-4, doi: <https://doi.org/10.1007/s11943-018-0234-z>.
4. Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N. (2018). **The R package emdi for estimating and mapping regionally disaggregated indicators**, accepted for the Journal of Statistical Software. Preliminary work was done in Kreutzmann (2016).
5. Rojas-Perilla, N., Kreutzmann, A.-K., and Medina, L. (2018). **A guideline of transformations in linear and linear mixed regression models**. *Working paper*, to be submitted. The work is an extension of Medina (2017).
6. Medina, L., Castro, P., Kreutzmann, A.-K., and Rojas-Perilla, N. (2018). **The R package trafo for transforming linear regression models**. *Working paper*, to be submitted.

Contents

Introduction	8
I Estimation of Disaggregated Linear Indicators in the Context of Household Surveys	10
1 The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany	11
1.1 Motivation	11
1.2 Data sources and initial analysis	13
1.2.1 The wealth survey: Panel on Household Finances	14
1.2.2 Register data of federal states and regional planning regions	17
1.3 Statistical method	18
1.3.1 The Fay-Herriot model	19
1.3.2 The log-transformed Fay-Herriot model	21
1.3.3 Combination of multiple imputation and the Fay-Herriot approach	22
1.3.4 Benchmarking for internal consistency	23
1.4 Application	23
1.4.1 Model selection and diagnostic checking	23
1.4.2 Gain in accuracy	25
1.4.3 Benchmarking	25
1.5 Discussion of the estimation results	26
1.6 Conclusion	28
Supplementary material A	30
2 The fayherriot command for estimating small-area indicators	35
2.1 Introduction	35
2.2 The Fay-Herriot model	36
2.2.1 Modeling	36
2.2.2 Estimating the variance of the random error	37
2.2.3 Evaluating the precision	37
2.2.4 Dealing with model-assumption violations	38
2.2.5 Overview of functionalities	38
2.3 The fayherriot command	40

2.3.1	Syntax	40
2.3.2	Options for fayherriot	40
2.3.3	predict after fayherriot:	41
2.3.4	Stored results	42
2.4	Example	42
2.4.1	Data description and direct estimates	42
2.4.2	Estimation using fayherriot	43
2.4.3	Comparison of direct and FH estimates	46
2.5	Conclusion	47
2.6	Acknowledgments	47
 II Estimation of Disaggregated Non-Linear Indicators based on Income and Wealth Data		48
 3 Estimation of sample quantiles: Challenges and issues in the context of income and wealth distributions		49
3.1	Motivation	49
3.2	Quantile definitions in statistical software	52
3.2.1	The default options and suggestions from the literature	52
3.2.2	Incorporation of sampling weights	55
3.2.3	Variance estimation	56
3.3	Comparison of quantile definitions	58
3.3.1	Simulation using income type data	59
3.3.2	Simulation using wealth data	63
3.4	Conclusion	67
3.5	Supplementary material	68
 Appendix B		69
B.1	Description of the Appendix	69
B.1.1	Inverse of the empirical cumulative distribution function	69
B.1.2	Observation closest to np	69
B.1.3	Linear interpolation of the empirical distribution function	70
B.1.4	Approximation to $F(E(X_k))$ for the normal distribution	70
B.1.5	Six desirable properties for sample quantile	70
 Supplementary material B		71
 4 The R package emdi for the estimation and mapping of regional disaggregated indicators		83
4.1	Introduction	83
4.2	Statistical methodology	85
4.2.1	Direct estimation	86
4.2.2	Model-based estimation	88

4.3	Data sets	90
4.4	Basic design and core functionality	92
4.4.1	Estimation of domain indicators	94
4.4.2	Summary statistics and model diagnostics	96
4.4.3	Selection and comparison of indicators	98
4.4.4	Mapping of the estimates	102
4.4.5	Exporting the results	103
4.5	Additional features	104
4.5.1	Incorporating an external indicator	104
4.5.2	Parallelization	105
4.6	Conclusion and future developments	107
Appendix C		109
C.1	Semi-parametric wild bootstrap	109
C.2	Reproducibility	110
 III Transformations for Achieving Model Assumptions in Linear and Linear Mixed Models		 112
5	A guideline of transformations in linear and linear mixed regression models	113
5.1	Introduction	113
5.2	Transformations step framework	115
5.2.1	Choose the model and be aware of the corresponding assumptions . . .	115
5.2.2	Choose a suitable transformation that addresses assumption violations .	116
5.2.3	Parameter inference and interpretation	139
5.3	Further issues with regards to variable transformations	141
5.4	Conclusions and future research directions	147
6	The R package trafo for transforming linear regression models	148
6.1	Introduction	148
6.2	Transformations	149
6.3	Study case	152
6.3.1	Finding a suitable transformation	153
6.3.2	Comparing the untransformed model with a transformed model	155
6.3.3	Comparing two transformed models	156
6.4	Customized transformation	159
6.5	Introduction	160
Appendix D		161
D.1	Likelihood derivation of the transformations	161
D.1.1	Log (shift) transformation	161
D.1.2	Glog transformation	162
D.1.3	Neglog transformation	163

D.1.4	Reciprocal transformation	164
D.1.5	Box-Cox (shift) transformation	165
D.1.6	Log-shift opt transformation	166
D.1.7	Bickel-Docksum transformation	167
D.1.8	Yeo-Johnson transformation	168
D.1.9	Square root-shift opt transformation	170
D.1.10	Manly transformation	171
D.1.11	Modulus transformation	172
D.1.12	Dual power transformation	173
D.1.13	Gpower transformation	175
Bibliography		177
Summaries		208
	Abstracts in English	208
	Kurzzusammenfassungen auf Deutsch	210

Introduction

International institutions and national statistical institutes are increasingly expected to report disaggregated indicators, i.e., means, ratios or Gini coefficients (Gini, 1912) for different regional levels, socio-demographic groups or other subpopulations (Piacentini, 2014; Leadership Council of the Sustainable Development Solutions Network, 2015). These subpopulations are called areas or domains in this thesis. The data sources that are used to estimate these disaggregated indicators are mostly national surveys which may have small sample sizes for the domains of interest. Therefore, direct estimates that are based only on the survey data might be unreliable. To overcome this problem, small area estimation (SAE) methods help to increase the precision of survey-based estimates without demanding larger and more costly surveys (Rao and Molina, 2015). In SAE, the collected survey data is combined with other data sources, e.g., administrative and register data or data that is a by-product of digital activities (Marchetti et al., 2015; Schmid et al., 2017).

The data requirements for various SAE methods depend to a large extent on whether the indicator of interest is a linear or non-linear function of a quantitative variable. For the estimation of linear indicators, e.g., the mean, aggregated data is sufficient, that is, direct estimates and auxiliary information from other data sources only need to be available for each domain. One popular area-level approach in this context is the Fay-Herriot model (Fay and Herriot, 1979) that is studied in Part I of this work. In Chapter 1, the Fay-Herriot model is used to estimate the regional distribution of the mean household net wealth in Germany. The analysis is based on the Household Finance and Consumption Survey (HFCS) that was launched by the European Central bank and several statistical institutes in 2010 (Household Finance and Consumption Network, 2016b). The main challenge of applying the Fay-Herriot approach in this context is to handle the issues arising from the data: a) the skewness of the wealth distribution, b) informative weights due to, among others, unit non-response, and c) multiple imputation to deal with item non-response (Rubin, 1987). For the latter, a modified Fay-Herriot model that accounts for the additional uncertainty due to multiple imputation is proposed in this thesis. It is combined with known solutions for the other two issues and applied to estimate mean net wealth at low regional levels. The Deutsche Bundesbank that is responsible for reporting the wealth distribution in Germany, as well as many economic institutes, predominantly work with the statistical software **Stata** (StataCorp, 2015). For providing the Fay-Herriot model and its extensions as used in Chapter 1, a new **Stata** command called `fayherriot` is programmed in the context of this thesis to make the approach available for practitioners. Chapter 2 describes the functionality of the command with an application to income data from the Socio-Economic Panel, one of the largest panel surveys in Germany (Goebel et al., 2018). The example appli-

cation demonstrates how the Fay-Herriot approach helps to increase the reliability of estimates for mean household income compared to direct estimates at three different regional levels.

In an extension to estimating linear indicators, Part II deals with the estimation of non-linear income and wealth indicators. Since the mean is sensitive to outliers, the median and other quantiles are also of interest when estimating the income or wealth distribution. As a first approach, this thesis focuses on the direct estimation of quantiles, which is not as straightforward as for the mean. In Chapter 3, common quantile definitions implemented in standard statistical software are empirically evaluated based on income and wealth distributions with regards to their bias. The analysis shows that, especially for wealth data that is mostly heavily skewed, sample sizes need to be large in order to obtain unbiased direct estimates with the common quantile definitions. Since a design-unbiased direct estimator is one assumption of the aforementioned Fay-Herriot model, further research would be necessary in order to use the Fay-Herriot approach for the estimation of quantiles when the underlying data is heavily skewed. More common methods for producing reliable estimates for non-linear indicators – including quantiles, poverty indicators (Foster et al., 1984), and inequality indicators such as the Gini coefficient (Gini, 1912) – in small domains are unit-level SAE methods. However, for these methods, the data requirements are more restrictive. Both the survey data and the auxiliary data need to be available for each unit in each domain. Among others, the empirical best prediction (EBP) (Molina and Rao, 2010), the World-Bank method (Elbers et al., 2003), and the M-Quantile approach (Chambers and Tzavidis, 2006) are well-known methods for the estimation of non-linear indicators in small domains. However, these methods are either not available in statistical software or the user-friendliness is limited. Therefore, in this work the R package **emdi** is developed that focuses on an user-friendly application of the EBP. Chapter 4 describes how the package **emdi** supports the user beyond the estimation by tools for assessing and presenting the results.

Both, area- and unit-level SAE models, are based on linear mixed regression models that rely on a set of assumptions, particularly the linearity and normality of the error terms. If these assumptions are not fulfilled, transforming the response variable is one possible solution. Therefore, Part III provides a guideline for the usage of transformations. Chapter 5 gives an extensive overview of different transformations applicable in linear and linear mixed regression models and discusses practical challenges. The implementation of various transformations and estimation methods for transformation parameters are provided by the R package **trafo** that is described in Chapter 6.

Altogether, this work contributes to the literature by

- a) combining SAE and multiple imputation proposing a modified Fay-Herriot approach,
- b) showing limitations of existing quantile definitions with regards to the bias when data is skewed and the sample size is small,
- c) closing the gap between academic research and practical applications by providing user-friendly software for the estimation of linear and non-linear indicators, and
- d) giving a framework for the usage of transformations in linear and linear mixed regression models.

Part I

Estimation of Disaggregated Linear Indicators in the Context of Household Surveys

Chapter 1

The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany

1.1 Motivation

The financial crisis that began in 2007 has uncovered the fragility of the global financial system. Private households' solvency is considered as one of the most important channels affecting financial stability. The tight linkage between private households and financial institutions is reflected by the large share of an economy's wealth, which is held by individuals (Ampudia et al., 2016). For instance, total financial assets of private households in Germany accounted for almost € 6 trillion, which is almost twice as large as the German annual GDP of € 3.3 trillion in 2017 (Deutsche Bundesbank, 2018). A historical perspective undermines this important relationship as two thirds of worldwide financial crises were preceded by a mortgage lending boom (Brunnermeier and Schnabel, 2016).

Several scholars have underpinned the important impact of the distribution of income and wealth on the stability of the financial system. For the United States (US), Kumhof et al. (2015) show the relationship between increasing income inequality, wealth concentration and their impact on financial stability. The key mechanism lies in the negative marginal propensity to consume (see e.g., Carroll, 1998). An increasing income inequality results in a higher concentration of savings by top earners in form of loans to bottom earners. This may lead to a rise of the debt-to-income ratio, which in turn may pose threats to the stability of the financial system.

Most of the literature on the economic importance of income and wealth distributions is focused at the national distribution. In this context, the spatial distribution of economic activity deserves additional attention, as economic activities are unevenly distributed across space (see e.g., Ottaviano and Puga, 1998). These agglomeration economies can be attributed to the importance of the local concentration of wealth, economic activity and innovative capacity (see e.g., Rodríguez-Pose and Crescenzi, 2008), but also to a process of rising inequality across regions. This regional divergence provides the justification for financial support schemes such

as the European Cohesion Funds allocating large parts of the EU's total budget.

The rising importance of agglomeration economies with their linkage to private wealth provides the motivation to assess the regional distribution of private financial resources (eurostat, 2017). In this context, the German reunification process provides a compelling example from an economic point of view. After a rapid catching-up process driven mostly by the construction sector, the convergence of East German regions came to abrupt end in 1995. Afterwards the disposable income of households living in East Germany has been stagnating at about 80% of the West German level (Blum et al., 2010), while private net wealth in the East has caught up only to about 40% of the West German level (Deutsche Bundesbank, 2016).

Differences in the income and wealth gap between East and West Germany are rooted in the economic conditions after the reunification. The ratio of disposable income of East German to West German households was 46% in 1991, whereas East German mean net wealth corresponded only to 30% of the average net wealth of households living in West Germany (Ammermüller et al., 2005). Blum et al. (2010) provide several reasons behind the initial difference between the income and wealth. First, the weak economy of East Germany at the start of the reunification process translated into price differences of capital goods such as house prices. Second, the institutional setting in the former German Democratic Republic (GDR) including a low protection of property rights provided weaker incentives for private capital accumulation. Third, these differences in incentives for wealth accumulation translated into lower home-ownership rates and lower saving rates in East Germany. Fourth, financial assets were converted at the rate 2:1 in contrast to the 1:1 conversion rate of wages. Furthermore, a lower savings ratio in East Germany contributed to the persistence in the wealth gap.

In order to lower the income and wealth gap and to foster the convergence process of East German regions, the German government implemented several programs. The so-called *Funds for German Reunification* (in German: Fonds Deutsche Einheit) was succeeded by two programs labeled *Solidarity Pact* (in German: Solidarpakt). The *Solidarpakt II* will expire in 2019, and it is an ongoing debate whether a national scheme for regional cohesion shall focus exclusively on the distinction between East and West Germany (Blum et al., 2011). Structurally weaker regions in the West claim that they are equally entitled for the reception of financial help. However, little is known so far about the distribution of financial resources among smaller regions within the East and West due to limited data availability. This article may enrich this discussion by focusing on the distribution of private wealth and financial resources at a regional level going beyond the distinction between East and West.

For the regional distribution of household income on different levels several data sources are already available (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017; RWI;microm, 2016a). Especially, information of income received in form of tax declarations or by notification of social security payments can help to develop indicators of income for low regional levels. In contrast, the measurement of the regional distribution of private wealth is not that trivial because the taxation of wealth is effective only in a few countries. Thus, surveys may provide one approach to capture the distribution of private wealth. Since 2011, the Household Finance and Consumption Network (HFCN) provides the Household Finance and Consumption Survey (HFCS) which is a survey that is, so far, conducted in 15 euro countries in the first wave

and in 19 euro countries in the second wave. It contains detailed information, among others, of household wealth on a micro-level. The HFCN states the importance of private wealth as an indicator for consumer spendings and its relevance for financial stability (Household Finance and Consumption Network, 2016b). In Germany, the Deutsche Bundesbank is responsible to describe the wealth distribution and since 2011 they use the German part of the HFCS, the Panel on Household Finances (PHF) (Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank, 2014). However, the lowest regional level for which estimates are reported are four larger regions with a distinction between East and West Germany. The latter is divided in three subregions. Given the sample size of about 4,500 households, a more disaggregated consideration is subject to concerns about the precision of estimates. Therefore, this paper makes use of small area estimation (SAE) in order to obtain reliable estimates for private net wealth at a lower regional level, namely 16 federal states (in German: Bundesländer) and 96 regional planning regions (in German: Raumordnungsregionen). From an applied perspective, this paper contributes to the discussion of the wealth distribution at a regional level and its implications on the allocation of regional support schemes in Germany.

The estimates are obtained by the Fay-Herriot model (Fay and Herriot, 1979). In order to increase the accuracy of estimates at lower regional levels, direct estimates obtained from survey data are enriched with covariate information from other data sources like registers. The challenge of applying the Fay-Herriot model in this work is the consideration of the data structure while using the SAE approach. First, the skewness of the wealth distribution requires the usage of a log-transformation in the Fay-Herriot approach for the planning regions in order to fulfill the normality assumptions (Slud and Maiti, 2006; Neves et al., 2013). Second, the present unit and item non-response needs to be taken into account. The unit non-response is adjusted by the data provider using weighting procedures. The produced sampling weights are considered in the Fay-Herriot model by using the weighted direct estimator in the model. The item non-response in the HFCS is handled with multiple imputation (Rubin, 1987). Therefore, our estimates are obtained by using a combination of Rubin's rule and the Fay-Herriot approach. From a theoretical perspective, this leads to a modified (transformed) Fay-Herriot that accounts for the additional uncertainty due to the multiple imputation. Third, for the reporting institution the internal consistency of the regional estimates with the estimate obtained for the national level needs to be ensured by benchmarking the model-based estimates.

The paper is structured as follows. Section 1.2 describes the data sources that are used in this work, particularly the PHF and the data sources for the covariate information. Section 1.3 describes the statistical method. In Section 1.4 the application of the Fay-Herriot model for the estimation of household net wealth is described. The results are interpreted in Section 1.5. Section 1.6 discusses further potential research.

1.2 Data sources and initial analysis

In this section, the definition of household (HH) net wealth is introduced and the data sources used in the analysis are described. Wealth is composed of several assets and liabilities. It can be measured as gross wealth, the sum of assets, or as net wealth, the difference between assets and

liabilities. Thus, negative net wealth is possible if liabilities exceed assets. A typical balance sheet of a HH is presented in the supplementary material of this paper. In order to take into account the HHs debt in this analysis, the HH net wealth is the variable of interest.

1.2.1 The wealth survey: Panel on Household Finances

Since wealth is an important indicator for financial stability, the central banks of the Eurosystem and several National Statistical Institutes initiated a joint survey called Eurosystem Household Finance and Consumption Survey (HFCS) as a consequence of the financial crisis in 2007. The survey provides detailed data on various aspects of HH balance sheets and related economic and demographic variables, including income, private pensions, employment and measures of consumption (Eurosystem Household Finance and Consumption Network, 2013a,b). The HFCS is the first harmonized survey data across eurozone countries and thus it is unique in enabling cross-country comparisons on a micro-level. Therefore, many studies are already based on this data. For instance, some studies compare the accumulated results for countries wealth from this micro-data source with macro-data sources like national accounts (Kavonius and Honkkila, 2013; Andreasch and Lindner, 2016).

The German part of the HFCS, namely the Panel on Household Finances (PHF) (Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank, 2014), is the data source that the Deutsche Bundesbank, the institution responsible to describe the wealth distribution in Germany, uses. An advantage of the PHF, compared to other data sources that cover wealth as the German Socio-Economic Panel (SOEP), is the detailed questioning of wealth components. However, the PHF bears some methodological issues that need to be taken into account for the analysis. The PHF is a panel survey with a first wave in 2011 and a second wave in 2014. While most of the following issues occur in both waves, all following numbers and the analysis itself is based on the second wave.

The sampling design aims to overrepresent wealthy households (Knerr et al., 2015). This is done because of the unequal distribution of assets, especially financial assets, and liabilities across households. The sampling is conducted in three stages. In the first stage, German municipalities are divided into three strata depending on the size and proportion of wealthy HHs. In a second stage, the streets in cities with more than 100,000 citizens are categorized in wealthy and other streets. In the third stage, the public register is used to draw persons above 18. This leads to a sample with 4,461 observations. The PHF has, however, a high unit non-response rate. Only 19% of the selected households participated in the survey. Thus, the data provider uses weighting procedures to adjust for the potential bias that is caused by the mentioned issues. A detailed description of the weighting procedure can be found in Knerr et al. (2015). Considering the sensitive questions, the low response rate is not surprising. In contrast, the HHs that decided to participate showed high item response rates (Eisele and Zhu, 2013). Thus, the item non-response is relatively low for many core variables even though very sensitive financial questions are asked. Nevertheless, the missing values due to item non-response have to be taken into consideration. Therefore, the institutions responsible for the survey, in Germany the Deutsche Bundesbank, are required to conduct multiple imputation (MI) according to Rubin (1987) to replace missing values (Household Finance and Consumption Network, 2016a).

Table 1.1: Pearson correlation coefficients of the variable net wealth in the five imputed data sets M1-M5 based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

	M1	M2	M3	M4	M5
M1	1.000	0.988	0.990	0.989	0.992
M2	0.988	1.000	0.995	0.993	0.994
M3	0.990	0.995	1.000	0.993	0.994
M4	0.989	0.993	0.993	1.000	0.993
M5	0.992	0.994	0.994	0.993	1.000

Thus, the imputation is conducted $M = 5$ times by the data provider that has more available information for the imputation model than the data user (Eisele and Zhu, 2013). For a proper imputation, Rubin (1996) suggests to use as many variables as available, especially when data provider and data analyst are distinct entities. The imputation in the PHF generally follows this rule. However, Eisele and Zhu (2013) also describe how they avoid overfitting using cross-validation methods for the variable selection. Besides variables that are correlated with the imputed variable, characteristics explaining the non-response behavior and design weights are included in the specifications. Disregarding the design could lead to a bias in the variance (Kott, 1995). Also domain indicators are considered (Household Finance and Consumption Network, 2016a). Since the item non-response is relatively low for many variables and especially for variables with a high impact on wealth, the difference between the five imputations of the variable net wealth is rather small which leads to a high correlation between the imputations (see Table 1.1). In order to receive final estimates in the presence of multiple data sets the estimates based on each data set need to be pooled. For the pooling, Rubin (1987) suggests a rule that is explained more detailed in Section 1.3.3. For the variance estimation of linear and non-linear indicators replication weights can be used in the PHF. These replication weights are received from a Rao-Wu rescaled bootstrap (Rao and Wu, 1988; Household Finance and Consumption Network, 2016a) and provided with the data.

Considering these issues, the mean HH net wealth is reported by the Deutsche Bundesbank for the regions East and West and the West is further divided into three subregions of northern, western and southern states (Deutsche Bundesbank, 2016). Table 1.2 summarizes the average net wealth level for these regions in thousand euro (TEUR) visualizing the distinct differences between East and West Germany. Furthermore, it provides an indication for heterogeneity of mean net wealth within the western part of Germany. Since our analysis is based on the most recent release of the scientific use file issued in 2017, the reported values differ slightly from those reported by Deutsche Bundesbank (2016).

This regional division (northern, western, southern and eastern states) is neither based on administrative units nor does the analysis highlight differences within the East. Since the Solidarpakt determines money transfers between the federal states, the wealth levels should also be estimated on the federal state level (BL). Furthermore, the federal states have own parliaments and budgets, hence identifying regions within the states with different wealth levels is also of interest. The federal states consist of regional planning regions that are used e.g., for

Table 1.2: Mean of HH net wealth in TEUR and sample sizes in the East, West and northern, southern and western federal states based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

Regional level		HH net wealth	Sample size
West		248.48	3610
	Northern states	Schleswig-Holstein, Hamburg, Niedersachsen, Bremen	256.66 752
	Southern states	Hessen, Baden-Württemberg, Bayern	285.32 1714
	Western states	Nordrhein-Westfalen, Rheinland- Pfalz, Saarland	196.83 1144
East		Thüringen, Sachsen, Sachsen- Anhalt, Brandenburg, Berlin, Mecklenburg-Vorpommern	90.23 851

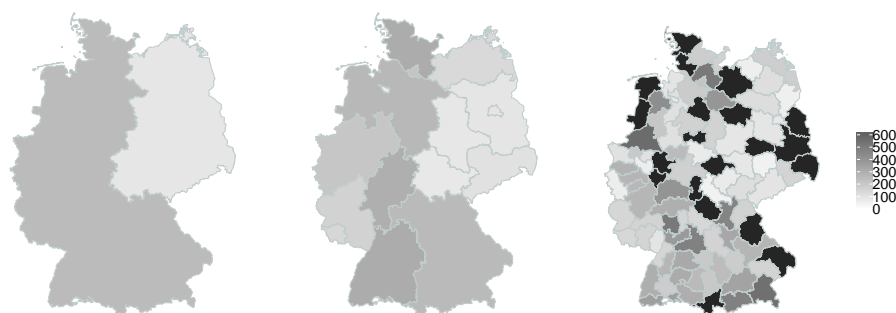


Figure 1.1: Map of the direct estimates of mean HH net wealth in TEUR for the regions East and West, the federal states and planning regions based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations. Out-of-sample regions are colored in black.

calculations of urban development promotion or housing benefits (Arndt et al., 2009). Der Paritätische Gesamtverband (2017) uses this regional disaggregation for the analysis of poverty. The estimation of indicators for the regional planning regions allows to investigate differences between urban and rural areas. Therefore, average HH net wealth is also estimated for the 96 planning regions (ROR) in this work. Preliminary results using direct estimation are shown in Figure 1.1. While the map of estimates for the East and the West only shows the known wealth difference between these two regions, the maps for the federal states and especially the estimates for the planning regions confirm the assumption of noticeable regional differences. However, Table 1.3 also shows a decrease in sample sizes compared to the sample sizes in the Table 1.2. Furthermore, the black colored planning regions do not have observations to directly estimate HH net wealth. Ten planning regions are not observed in the sample. For another nine regions the sample size is too small to obtain results not violating confidentiality issues by the Bundesbank. This means that the direct estimator would be only available for 77 out of 96 regions.

Thus, the application of a direct estimator for the mean of HH net wealth bears two issues:

Table 1.3: Summary of sample sizes in the federal states and planning regions based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

	Min.	1st Qu.	Median	3rd Qu.	Max.
BL	32	98	189	358	925
ROR	9	28	40	65	340

First, the direct estimates might be unreliable due to large variances in small areas. Second, direct estimates cannot be reported for regions with zero sample size or for direct estimates violating confidentiality issues. The combination of the survey data with register data that is available on these regional levels can help to improve the accuracy of the estimates for both regional levels and to provide predictions for the planning regions where direct estimation is not possible or not allowed to be published.

1.2.2 Register data of federal states and regional planning regions

The model used in this work requires additional information from administrative data sources that helps to predict the target variable HH net wealth. As already mentioned, net wealth is composed of several assets and liabilities. Thus, these components are natural predictors for the target variable. The HFCN states that home-ownership and the value of the real estate is the component with an especially strong influence on the HHs wealth (Eurosystem Household Finance and Consumption Network, 2013b; Household Finance and Consumption Network, 2016b). But also information of other assets like the number or value of vehicles or information of liabilities may be good predictors.

While not a component of net wealth, a higher income or income growth enables to accumulate wealth according to the income-to-wealth ratio proposed by Piketty and Zucman (2014). Since labor income is the highest proportion of the HH income for an average HH, employment figures such as the employment status should be taken into consideration. Especially, self-employed people are a group that tends to have higher wealth (Frick and Grabka, 2009). It can also be shown that net wealth initially increases with age and declines after retirement (Eurosystem Household Finance and Consumption Network, 2013b; Household Finance and Consumption Network, 2016b). Since we measure HH net wealth instead of personal net wealth, the HH structure can also have an effect on the level of net wealth. Single HHs tend to have a lower wealth but for HHs with two or more members the wealth does not increase with size (Eurosystem Household Finance and Consumption Network, 2013b). The variables that we use to proxy these effects are summarized in Table 1.4. It can be noted that the register information is only available on an aggregated level, i.e., this information is not available for each HH but for each region. This is mostly due to confidentiality issues which is an important issue especially in developed countries like Germany.

For the federal states most of these variables are obtained from the Federal Office of Statistics and the Statistical Offices of the Länder (Statistische Ämter des Bundes und der Länder, 2018) that maintain different databases containing regional information. For instance, the unemployment rate is obtained from the Regionaldatenbank Deutschland and the disposable in-

Table 1.4: Identified variables that potentially help to predict HH net wealth. The numbers in parenthesis state the number of variables for this group. References to the sources of the variables are given in the supplementary material.

Influence	Variables	Year	Level
Real estate	Ownership rate	2011	BL, ROR
	Rental and purchase prices per sqm, level of interior (3)	2014	BL, ROR
	Number of houses (2) and types of buildings (6)	2015	ROR
Vehicles	Density of cars and car segment (11)	2015	ROR
Savings	Saving quota of HH	2014	BL
Liabilities	Default probability (8)	2015	ROR
	Private debtors per 100 inhabitants	2014	ROR
Income	Disposable income of private HHs per inhabitant	2014	BL
	Average HH income per inhabitant	2014	ROR
Employment status	Unemployment rate,	2014	BL, ROR
	percentage of employees, civil servants and self-employed	2011	BL
Age	Age groups (4)	2014	BL, ROR
	Youth dependency ratio,	2014	BL
	old-age dependency ratio	2014	BL
Household structure	Single or couple	2015	ROR

come is received from the national accounts of the federal states. The covariate information for the planning regions is predominantly provided by the research data center FDZ Ruhr - RWI (Budde and Eilers, 2014; RWI;microm, 2016a) and complemented by the German database Indikatoren und Karten zur Raum- und Stadtentwicklung (INKAR) (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017). Information about the rental and purchase prices as well as the level of interior is delivered by the empirica ag for both regional levels (Empirica ag, 2017). Most variables are obtained for 2014 which is the year of the survey or for 2015. We assume that these variables are relatively time consistent on the aggregated level. For the same reason, we include the variables about the type of employment and ownership rates even though these are obtained from the German Census in 2011 (Statistische Ämter des Bundes und der Länder, 2011). Other factors that have an essential influence on wealth accumulation are inheritances and donations. However, data for these variables was not available for this work.

1.3 Statistical method

In this section, the statistical methodology for receiving estimates for the mean of HH net wealth for the German federal states and planning regions is described. With regard to the

methodological issues mentioned in Section 1.2.1 and the requirements of institutions that provide official statistics, the method for the desired application should

- help to increase the accuracy of the direct estimates,
- help to receive estimates for domains with a sample size of zero or confidentiality issues,
- be able to handle the complex survey design and the uncertainty due to MI,
- should return estimates that are consistent with the direct estimates of the regions East and West and the national direct estimate.

Therefore, we propose a benchmarked Fay-Herriot (FH) estimator (Fay and Herriot, 1979) that additionally accounts for the variability due to the MI. The FH approach is one out of a wide range of SAE methods that generally combine information from different data sources. For an overview of SAE methods we refer to Pfeffermann (2013), Rao and Molina (2015) and Tzavidis et al. (2018). The main benefit of the FH model compared to other SAE methods is that it only requires additional information on an aggregated level. This is especially useful in Germany due to strict data protection rules. The following section explains the method in detail.

1.3.1 The Fay-Herriot model

In SAE, following setup is generally assumed. A finite population of size N is partitioned into D domains of sizes N_1, \dots, N_D , where $d = 1, \dots, D$ refers to a d th domain and $i = 1, \dots, N_d$ to the i th HH/individual. A sample is drawn from this population using a complex sampling design with sample sizes n_1, \dots, n_D .

The FH model assumes two model relations. The sampling model can be expressed as

$$\hat{\theta}_d^{\text{Dir}} = \theta_d + e_d, \quad d = 1, \dots, D, \quad (1.1)$$

where $\hat{\theta}_d^{\text{Dir}}$ is a design-unbiased direct estimator of the population indicator θ_d , for example a mean. It is assumed to be equal to the population value, θ_d , plus a sampling error e_d . The direct estimator allows the incorporation of sampling weights w and thus the requirement of handling the survey design is fulfilled. In this work, the indicator of interest is the mean of HH net wealth. The weighted mean for each domain d is defined as follows

$$\hat{\theta}_d^{\text{Dir}} = \frac{\sum_{i=1}^{n_d} w_{di} y_{di}}{\sum_{i=1}^{n_d} w_{di}}, \quad i = 1, \dots, n_d, \quad d = 1, \dots, D,$$

where y_{di} is the target variable and w_{di} are the sampling weights for domain d and HH i .

The second model links the population indicator θ_d with covariate information \mathbf{x} in a linear relation as follows

$$\theta_d = \mathbf{x}_d^\top \boldsymbol{\beta} + u_d, \quad d = 1, \dots, D, \quad (1.2)$$

where \mathbf{x}_d is a $p \times 1$ vector of area-level covariate information and $\boldsymbol{\beta}$ is the vector of regression parameters with dimension $p \times 1$. The combination of the models 1.1 and 1.2 leads to a special linear mixed model that is defined as

$$\begin{aligned}\hat{\theta}_d^{\text{Dir}} &= \mathbf{x}_d^\top \boldsymbol{\beta} + u_d + e_d, \quad d = 1, \dots, D, \\ u_d &\stackrel{iid}{\sim} N(0, \sigma_u^2) \quad e_d \stackrel{iid}{\sim} N(0, \sigma_{e_d}^2),\end{aligned}$$

with random effects u_d that are independent and identically normally distributed with variance σ_u^2 and sampling errors e_d that are independent normally distributed with variance $\sigma_{e_d}^2$. The two error terms are assumed to be independent. The estimates of the regression parameters $\hat{\boldsymbol{\beta}}$ are the empirical best linear unbiased estimators (EBLUE) of $\boldsymbol{\beta}$ (Rao and Molina, 2015). For the estimation of the variance of the random effect σ_u^2 , several approaches are available, among others, the FH method-of-moments estimator, the maximum likelihood (ML) and the residual maximum likelihood (REML) method (Rao and Molina, 2015). A disadvantage of these approaches is the numerical possibility of a negative variance estimator that is usually set to 0. This issue may especially arise in the case of a small number of domains. Therefore, adjusted estimation methods can be preferable when the number of domains is small since these always provide strictly positive variance estimators (Li and Lahiri, 2010; Yoshimori and Lahiri, 2014). Yoshimori and Lahiri (2014) propose an adjusted maximum residual likelihood approach (AMRL.YL).

The resulting FH estimator is an empirical best linear unbiased predictor (EBLUP) of θ_d . It can be expressed as a weighted average of the direct estimator $\hat{\theta}_d^{\text{Dir}}$ and a synthetic part as follows

$$\begin{aligned}\hat{\theta}_d^{\text{FH}} &= \mathbf{x}_d^\top \hat{\boldsymbol{\beta}} + \hat{u}_d \\ &= \hat{\gamma}_d \hat{\theta}_d^{\text{Dir}} + (1 - \hat{\gamma}_d) \mathbf{x}_d^\top \hat{\boldsymbol{\beta}},\end{aligned}$$

where $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_d}^2}$ is the ratio of the variance of the random effects and the total variance and denotes the shrinkage factor for area d . Whenever the variance of the sampling error is relatively small, $\hat{\gamma}_d$ gets large and more weight lies on the direct estimator. This feature of the FH model is especially desired in official statistics for justifying a model-based approach since it uses the direct estimator when it is reasonable. In the other case, when the sampling error variance is relatively large, less weight lies on the direct estimator. For the domains with zero sample size or with estimates that are not allowed to be published the prediction shrinks to the synthetic part

$$\hat{\theta}_{d,\text{out}}^{\text{FH}} = \mathbf{x}_d^\top \hat{\boldsymbol{\beta}}.$$

In order to assess the accuracy of the FH estimates the corresponding mean squared error (MSE) can be estimated. The composition of the MSE estimator depends on the selected estimation method for the variance of the random effect (Rao and Molina, 2015). Prasad and Rao (1990) and Datta and Lahiri (2000) describe the compositions when the REML and ML approaches are used, respectively. The MSE estimation of out-of-sample domains for both approaches follows Rao and Molina (2015). For the adjusted estimation methods, the MSE

needs to be modified as described in Li and Lahiri (2010) and Yoshimori and Lahiri (2014).

1.3.2 The log-transformed Fay-Herriot model

If the relationship between the target variable and the covariate information is nonlinear or the normality assumption of the error terms is not met the log-transformed FH model can be used (Rao, 1999). According to Neves et al. (2013), the direct estimator and the sampling error variance can be transformed as follows.

$$\begin{aligned}\hat{\theta}_d^{\text{Dir}^*} &= \log(\hat{\theta}_d^{\text{Dir}}), \\ \text{var}(\hat{\theta}_d^{\text{Dir}^*}) &= \left(\hat{\theta}_d^{\text{Dir}}\right)^{-2} \text{var}(\hat{\theta}_d^{\text{Dir}}),\end{aligned}$$

where the * denotes the transformed scale. The FH estimator on the transformed scale can be obtained by using the log-transformed direct estimator $\hat{\theta}_d^{\text{Dir}^*}$ as dependent variable and the modified variance $\text{var}(\hat{\theta}_d^{\text{Dir}^*})$ as estimate for the sampling error variance.

$$\hat{\theta}_d^{\text{FH}^*} = \hat{\gamma}_d^* \hat{\theta}_d^{\text{Dir}^*} + (1 - \hat{\gamma}_d^*) \mathbf{x}_d^\top \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_d^* = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_d,*}^2}$ with $\sigma_{e_d,*}^2 = \text{var}(\hat{\theta}_d^{\text{Dir}^*})$. For an appropriate interpretation of the results, the FH estimators need to be back-transformed to the original scale. A naive back-transformation using the exponential function may induce a bias to the estimates because of Jensen's inequality (Jensen, 1906). Therefore, several bias-corrected back-transformations are proposed in the literature. Neves et al. (2013) back-transform $\hat{\theta}_d^{\text{FH}^*}$ based on the properties of the log-normal distribution which we refer to as crude back-transformation:

$$\begin{aligned}\hat{\theta}_d^{\text{FH, crude}} &= \exp\left\{\hat{\theta}_d^{\text{FH}^*} + 0.5 \text{MSE}(\hat{\theta}_d^{\text{FH}^*})\right\}, \\ \text{MSE}(\hat{\theta}_d^{\text{FH, crude}}) &= \exp\left\{\hat{\theta}_d^{\text{FH}^*}\right\}^2 \text{MSE}(\hat{\theta}_d^{\text{FH}^*}),\end{aligned}$$

where $\hat{\theta}_d^{\text{FH}^*}$ is the FH estimator on the transformed scale and $\text{MSE}(\hat{\theta}_d^{\text{FH}^*})$ the MSE estimator on the transformed scale, e.g., the Prasad-Rao MSE.

Slud and Maiti (2006) propose a bias-correction that considers the area-specific effects when the ML approach is used for the estimation of σ_u^2 . Chandra et al. (2018) extend this work by further taking into account the variability due to parameter estimation. The back-transformation for the point estimates differs slightly to the crude back-transformation:

$$\begin{aligned}\hat{\theta}_d^{\text{FH, Slud-Maiti}} &= \exp\left\{\hat{\theta}_d^{\text{FH}^*} + 0.5 \hat{\sigma}_u^2 (1 - \hat{\gamma}_d^*)\right\}, \\ \hat{\theta}_d^{\text{FH, Chandra et al.}} &= c_d * \exp\left\{\hat{\theta}_d^{\text{FH}^*} + 0.5 \hat{\sigma}_u^2 (1 - \hat{\gamma}_d^*)\right\},\end{aligned}$$

where c_d is a bias term derived in Chandra et al. (2018).

Furthermore, a special MSE estimator for the log-transformed model is developed in Slud and Maiti (2006). A disadvantage of the latter two bias-corrections is that these are only applicable for in-sample domains since these are based on the estimated $\hat{\gamma}_d^*$ which is only available for sampled domains.

1.3.3 Combination of multiple imputation and the Fay-Herriot approach

As already mentioned, the conducted MI needs to be considered. The imputation for the PHF is conducted by the data provider that is able to use more information about e.g., the non-response behavior. Information about the imputation is given in Section 1.2.1 and in Eisele and Zhu (2013). At this stage, we take the imputed data sets as appropriate and given. In the PHF, five values are imputed for each missing value which leads to five imputed data sets. The indicators of interest and its variances are estimated on each of these data sets. In order to pool these estimates, Rubin's rule can be applied when the complete data estimates are approximately normal (Rubin, 1987). Thus, it can be applied for the mean (Marshall et al., 2009). Consequently, we propose to use the direct estimator and the corresponding variance after the application of Rubin's rule defined as $\hat{\theta}_d^{\text{RRDir}}$ and $\hat{\sigma}_{\epsilon_d}^2 = \text{var}(\hat{\theta}_d^{\text{RRDir}})$ in the Fay-Herriot approach. The variance $\hat{\sigma}_{\epsilon_d}^2$ covers the sampling variance and the variance due to missingness and imputation (Rubin, 1996; Kim et al., 2006). The steps of the analysis are summarized as follows.

Step 1. Imputation: Impute the missing values. In the case of the PHF data set, the imputation is already conducted by the Deutsche Bundesbank (Eisele and Zhu, 2013).

Step 2. Analysis (Direct): Obtain $\hat{\theta}_{d,m}^{\text{Dir}}$ and $\text{var}(\hat{\theta}_{d,m}^{\text{Dir}})$ for $m = 1, \dots, M$ where M is the number of imputed data sets. For the PHF data set, M equals to 5.

Step 3. Pooling (Rubin's rule): Obtain $\hat{\theta}_d^{\text{RRDir}} = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_{d,m}^{\text{Dir}}$ and

$$\hat{\sigma}_{\epsilon_d}^2 = \frac{1}{M} \sum_{m=1}^M \text{var}(\hat{\theta}_{d,m}^{\text{Dir}}) + \left(1 + \frac{1}{M}\right) \frac{1}{M-1} \sum_{m=1}^M (\hat{\theta}_{d,m}^{\text{Dir}} - \hat{\theta}_d^{\text{RRDir}})^2.$$

Step 4. Analysis (FH): Obtain the Fay-Herriot estimator for multiple imputed data sets (FH-MI) expressed by:

$$\hat{\theta}_d^{\text{FH-MI}} = \hat{\gamma}_d \hat{\theta}_d^{\text{RRDir}} + (1 - \hat{\gamma}_d) \mathbf{x}_d^\top \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{\epsilon_d}^2}$.

Step 4*. Analysis (log-transformed FH): Log-transform the direct estimator and modify the variance estimator (here according to Neves et al. (2013)):

$$\begin{aligned} \hat{\theta}_d^{\text{RRDir}*} &= \log(\hat{\theta}_d^{\text{RRDir}}), \\ \hat{\sigma}_{\epsilon_d,*}^2 &= \left(\hat{\theta}_d^{\text{RRDir}}\right)^{-2} \hat{\sigma}_{\epsilon_d}^2. \end{aligned}$$

Obtain the FH estimator for multiple imputed data sets (FH-MI) on the transformed scale expressed by:

$$\hat{\theta}_d^{\text{FH-MI}*} = \hat{\gamma}_d^* \hat{\theta}_d^{\text{RRDir}*} + (1 - \hat{\gamma}_d^*) \mathbf{x}_d^\top \hat{\boldsymbol{\beta}},$$

where $\hat{\gamma}_d^* = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{\epsilon_d,*}^2}$.

Back-transform the estimation results to the original scale.

Step 5. Computation of the MSE: Obtain the MSE estimate for $\hat{\theta}_d^{FH-MI}$. As discussed in Section 1.3.1, the choice of the MSE estimator depends on the chosen estimation method for $\hat{\sigma}_u^2$ in Step 4.

Step 5*. Computation of the back-transformed MSE: Obtain the MSE estimate for the back-transformed $\hat{\theta}_d^{FH-MI*}$. The MSE estimator depends on the chosen bias-correction in Step 4*: crude, Slud-Maiti, Chandra et al. (see also Section 1.3.2).

1.3.4 Benchmarking for internal consistency

The aggregated regional FH-MI estimates can differ from the direct estimates on the national level or on the regional levels East and West. Thus, for meeting the requirement of internal consistency of the estimates, a benchmark approach following Datta et al. (2011) is utilized. This approach assumes the following relationship between the benchmark value τ and the aggregated values of the single regions.

$$\sum_{d=1}^D \xi_d \hat{\theta}_d^{FH-MI,bench} = \tau$$

where $\xi_d = \frac{N_d}{N}$ is the ratio of the population size in each region divided by the total population size. The benchmark value may be the national estimate or the estimate of larger regions like the regions East and West. The benchmarked FH-MI estimator can be expressed as

$$\hat{\theta}_d^{FH-MI,bench} = \hat{\theta}_d^{FH-MI} + \left(\sum_{d=1}^D \frac{\xi_d^2}{\phi_d} \right)^{-1} \left(\tau - \sum_{d=1}^D \xi_d \hat{\theta}_d^{FH-MI} \right) \frac{\xi_d}{\phi_d}.$$

Datta et al. (2011) present different ways to define ϕ_d . Some depend on the value of the estimate or on its accuracy estimate. If $\phi_d = \xi_d$, all FH-MI estimates are adjusted equally. When dividing ξ_d by the corresponding point or MSE estimate the domains with a respectively larger value are adjusted stronger. Steorts and Ghosh (2013) show that the MSE of the FH estimators increases only slightly because of the benchmarking. They propose a parametric bootstrap estimator for the estimation of the MSE of benchmarked estimates.

1.4 Application

In this section, the mean of HH net wealth is estimated for the German federal states and the planning regions using the FH method (Fay and Herriot, 1979) described in Section 1.3.

1.4.1 Model selection and diagnostic checking

The model presented in Section 1.3 depends on covariate information and several assumptions. Therefore, the model selection and some diagnostic checks are described before discussing the estimation results.

In Section 1.2.2, the predictors that can potentially explain HH net wealth and are available at the desired regional level were introduced. In total, the covariates sum up to 18 for the fed-

eral states and to 46 for the planning regions. Since the number of possible predictors exceeds the number of federal states and is high for the planning regions a variable reduction via an elastic net is conducted. Following Zou and Hastie (2005), an elastic net reduces the number of variables by eliminating trivial variables and including whole groups of closely related variables. The variable selection is based on the Kullback symmetric divergence criterion (KICb2) proposed by Marhuenda et al. (2014) especially for the FH model. Finally, the model with the lowest value of KICb2 is chosen for our analysis (Marhuenda et al., 2014).

For the federal states, the final model includes two covariates, the saving quota and the youth dependency ratio, capturing the relation between adolescents up to 19 years of age and individuals aged 20 to 64, with positive effects on the mean of HH net wealth. Since the number of domains is small with 16 domains, the REML and the AMRL.YL method are considered. For both methods, the estimation of σ_u^2 is similar and far from 0. Thus, the REML approach and the Prasad-Rao MSE are used since no adjustment is needed to receive a positive variance estimate. The explanatory power of the selected model measured by the modified R^2 for FH models proposed by Lahiri and Suntornchost (2015) is 92%. The normality assumption of the two error terms in the FH model is assessed by the Shapiro-Wilk test and is not rejected for both error terms. For the random effect (RE) the p-value equals 0.06 and for the standardized realized residuals (RRES) it is 0.85.

For the planning regions, the normality assumption of the error terms does not hold in the original scale. Thus, a log-transformation on the direct estimates of the mean of HH net wealth is applied. The transformed variable is used as dependent variable in the variable selection. The final model includes four covariates, the number of houses that are not for businesses, purchase price per sqm, the percentage of vans and the percentage of HHs with a default probability below average. The effect of the variables is positive and the modified R^2 is 89%. The Shapiro-Wilk test supports the assumption of normally distributed error terms (RE: p = 0.45, std. RRES: p = 0.43) in the log-transformed FH model. Thus, the FH and log-transformed FH model is used for the estimation of the mean of HH net wealth for the federal states and the regional planning regions, respectively. The FH-MI estimates in the transformed scale for the planning regions are back-transformed to the original using the crude back-transformation. This choice is based on the fact that the crude back-transformation is applicable for in- and out-of-sample domains. Furthermore, the differences between the estimates for in-sample domains using the crude back-transformation and the back-transformations proposed by Slud and Maiti (2006) and Chandra et al. (2018) are quite small.

One possibility to assess the quality of the model-based estimates is the comparison with the direct estimates. Brown et al. (2001) propose a goodness-of-fit test for this assessment. The null hypothesis of the test assumes that the model-based estimates do not differ significantly from the direct estimates. The test statistic is defined as

$$W(\theta_d^{\text{FH-MI}}) = \sum_{d=1}^D \frac{(\theta_d^{\text{RRDir}} - \theta_d^{\text{FH-MI}})^2}{\text{var}(\theta_d^{\text{RRDir}}) + \text{MSE}(\theta_d^{\text{FH-MI}})},$$

where $\theta_d^{\text{FH-MI}}$ is the FH-MI estimator for the federal states and the back-transformed $\theta_d^{\text{FH-MI}*}$ estimator for the planning regions and $\text{MSE}(\theta_d^{\text{FH-MI}})$ are the corresponding MSE estimates. Note

Table 1.5: Results for the goodness-of-fit test according to Brown et al. (2001).

Level	W	df	p-value
BL	4.23	16	0.99
ROR	22.84	77	1

Table 1.6: Distribution of the MSE of mean HH net wealth across federal states (BL) and planning regions (ROR) based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

Level	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
BL	Direct	140.47	430.41	821.16	1770.58	1906.36	8968.47
	FH-MI	143.54	371.31	552.89	592.14	844.02	1169.21
ROR	Direct	86.63	941.53	3701.64	19765.09	9610.40	561717.00
	FH-MI	81.40	766.19	1434.42	2432.02	3121.08	13846.04

that only the estimates of in-sample domains are compared because direct estimates cannot be obtained for out-of-sample domains. The test statistic W is χ^2 -distributed with D degrees of freedom under the null hypothesis. The results of the test show that the null hypothesis, model-based estimates do not differ significantly from the direct estimates, cannot be rejected (see Table 1.5). According to Chandra et al. (2015), a useful diagnostic that measures the adequacy of the model is the correlation coefficient of the synthetic part of the FH-MI estimates and the direct estimates. For the federal states this correlation is 0.88 and for the planning regions 0.68. Both values are comparable or higher to the value in Chandra et al. (2015).

1.4.2 Gain in accuracy

The accuracy of the estimates is measured by the Prasad-Rao MSE for both, the federal states and the planning regions, since the REML approach is used for the estimation of σ_u^2 . The MSE for the planning regions is further back-transformed to the original scale using the crude bias-correction as described in Section 1.3.2. Table 1.6 shows that the FH-MI estimator is more accurate than the direct estimator for the mean of HH net wealth for the federal states and planning regions. The gain in accuracy is especially large for the planning regions. From these results we can conclude that the FH approach helps to receive more reliable results.

1.4.3 Benchmarking

For the internal consistency the benchmarking approach described in Section 1.3.4 is implemented. The direct estimates in the regions East and West are used as a benchmark. These estimates also almost sum up to the national estimate with a negligible difference of 0.11% and 0.20% for the federal states and planning regions, respectively. Thus, benchmarking to the direct estimates of the regions ensures the consistency with the national estimate. Table 1.7 shows that the aggregated mean of HH net wealth of the FH-MI estimates for the planning regions overestimates the regional direct estimate for the East and underestimates the corresponding estimate for the West. For the benchmarking specification, $\phi_d = \xi_d/\text{FH-MI}_d$ is

Table 1.7: Mean difference of aggregated FH-MI estimates to the regional direct estimates for East and West in TEUR based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

Level	Benchmark	Regional direct estimate	Aggregated estimate
BL	East	90.23	89.77
	West	248.48	230.02
ROR	East	90.23	108.41
	West	248.48	238.40

Table 1.8: Distribution of the mean HH net wealth across federal states (BL) and planning regions (ROR) in TEUR based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

Level	Estimate	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
BL	Direct	69.68	93.19	153.01	175.81	254.63	307.74
	FH-MI	74.83	87.23	155.27	159.79	223.51	283.91
	FH-MI, bench	75.87	87.67	167.73	170.11	241.44	306.69
ROR	Direct	51.45	112.92	178.40	220.68	294.17	621.41
	FH-MI	63.79	124.85	174.69	202.26	254.52	512.31
	FH-MI, bench	53.09	123.88	179.02	205.37	265.28	533.97

chosen in order to adjust regions with a large estimate stronger. The order/ranks of the regions with regard to the value of the mean of HH net wealth remains unchanged within the East and the West. After the application of the benchmarking approach, the aggregated estimates are equal to the regional direct estimates and thus also equal to the national estimate. Table 1.8 shows the distribution of the mean HH net wealth across the federal states and the planning regions for the direct, the FH-MI and the benchmarked FH-MI estimates. It can be seen that most of the benchmarked results are larger than the FH-MI estimates. This is due to the fact that adjusting the underestimation of the aggregated estimate of the West has the larger effect on the benchmarked estimates than adjusting the overestimation or slight underestimation of the aggregated estimates of the East (see Table 1.7). For the discussion in Section 1.5, the benchmarked FH-MI estimates are used since these fulfill the requirement to sum up to the regional and national direct estimates.

1.5 Discussion of the estimation results

While the former section discusses the results from a statistical perspective, this section describes the regional distribution of wealth in Germany and sets the results into relation with theoretical knowledge.

Figure 1.2 shows the regional distribution of benchmarked FH-MI estimates for the states and the planning regions. The map for the federal states shows the fairly known pattern of a clear cut at the former border between East and West. All federal states in the East report an average private net wealth of TEUR 90, which is more than 50% below of the national mean reaching from about TEUR 75 in Saxony-Anhalt to TEUR 110 in Brandenburg. As outlined by

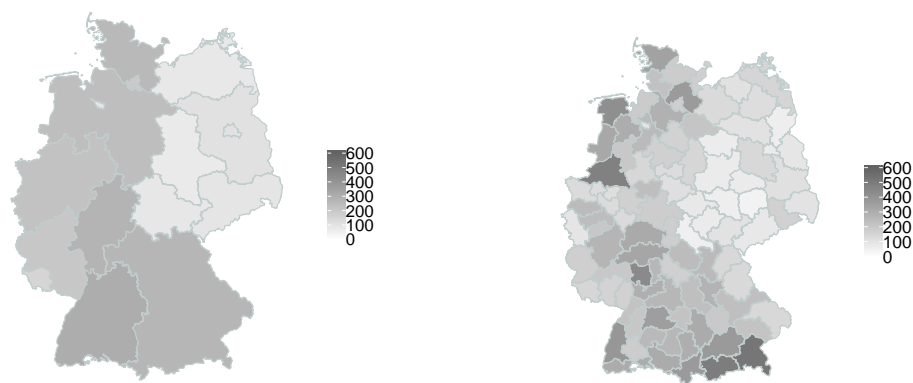


Figure 1.2: Map of the benchmarked FH-MI estimates for the federal states (left) and for the planning regions (right) of the mean of HH net wealth in TEUR based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations.

the Deutsche Bundesbank (2016), the strong separation with respect to private net wealth can be attributed to differences in financial wealth (TEUR 30 in the East to TEUR 60 in the West), home-ownership (35% to 47%) as well as in the average value of owned dwellings (TEUR 145 to TEUR 250). This relationship is in line with the ratio of average sqm-prices for dwelling provided by bulwiengesa AG (2018). In 2014, the average sqm-price for the purchase of a real estate in the East was below TEUR 1.2, whereas it was at TEUR 1.9 in the West.

Furthermore, the estimates also provide evidence for heterogeneity of private wealth across West German federal states. Our estimates confirm the findings by the Deutsche Bundesbank (2016) reporting that the average net wealth of federal states in the South (Baden-Württemberg, Bavaria and Hessen with an average of TEUR 284) is about 50% higher than in federal states located in the West (North Rhine-Westphalia, Rhineland-Palatinate, and Saarland with an average net wealth of TEUR 193). This difference is mostly driven by the conditional mean value of owner occupied housing (with TEUR 275 in the South and TEUR 197 in the West), financial assets (TEUR 73 to TEUR 53) and only slightly by home-ownership rates (48% to 44%). Furthermore, the federal city state of Berlin reports an average net wealth below TEUR 100. This observation is reasonable with regard to the home-ownership rates, which are often lower in the cities. This especially holds for Berlin with a home-ownership rate of 15.6% (Landesamt für Statistik Niedersachsen, 2014).

The analysis on the level of the planning regions enables further insights. Our results provide evidence for heterogeneity in West Germany. The regions around economically prosperous cities in the West – namely Munich, Frankfurt and Hamburg – report the highest private wealth levels in Germany. The top two regions (Südostoberbayern and Oberland) are located in the South of Munich, where average private net wealth is around TEUR 520. The other end of the distribution in the West predominantly contains regions in the Ruhr Area. This region was severely affected by the breakdown of the coal and steel industry, which is still reflected in the highest unemployment rates among West German planning regions. Note that the four planning regions of the Ruhr Area are listed among the 5 regions with highest unemployment

rates in West Germany: 1. Emscher-Lippe (11.7%), 2. Dortmund (11.5%), 3. *Bremen* (10.4%), 4. Duisburg/Essen (10.3%), and 5. Bochum/Hagen (9.1%) (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017). The results show that average private net wealth in these four planning regions are quite similar reaching from TEUR 115 in Dortmund to TEUR 125 in Duisburg/Essen.

The consideration of the 22 planning regions in East Germany including Berlin also captures regional differences of the distribution of private wealth. The four regions with the lowest estimates for private wealth in East Germany are geographically dispersed across four different federal states: Westsachsen in Saxony (TEUR 53), Südthüringen in Thuringia (TEUR 62), Uckermark-Barmin in Brandenburg (TEUR 66) and Halle/Saale in Saxony-Anhalt (TEUR 66). The regions with highest private wealth are located in the South-West of Berlin (Havelland-Fläming: TEUR 130), at the Baltic Sea (Vorpommern: TEUR 143) as well as in the region around the city of Dresden (Oberes Elbtal/Osterzgebirge: TEUR 154).

The results for German planning regions show that wealth is geographically dispersed in both parts of the country. Furthermore, we can show that private wealth in all East German planning regions still remains far below the national average. However, the wealthiest planning regions in the East report higher private wealth figures than the West German regions with lowest private wealth estimates.

1.6 Conclusion

The concentration of private income and wealth and the presence of financial support schemes among countries and regions motivate the assessment of the regional distribution of financial resources. While data sources for the estimation of regional income indicators are comprehensive, the current best source for the estimation of private wealth is, in most countries, survey data. In this context, the European Central Bank launched the HFCS in 2010, which is conducted in each country of the euro area. While the HFCS is, so far, used to report national estimates for private wealth, this work shows how to estimate average HH net wealth for low regional levels, namely the 16 federal states and 96 planning regions in Germany. We contribute to the literature by estimating the regional distribution of private wealth in Germany by means of a modified FH model, which

- a) accounts for the skewness of the wealth distribution by means of a log-transformation in the estimation,
- b) accounts for multiple imputation, and
- c) ensures internal consistency of the estimates with a national benchmark.

The results of the estimation are very insightful and contribute to the discussion on the distribution of private wealth, which has strikingly gained attention in the scientific literature as well as in the public debate in recent years. Even 25 years after the German reunification, there is a clear cut at the former border with respect to private wealth. However, the wealthiest planning regions in the East report higher private wealth figures than the West German regions with lowest private wealth estimates. This important finding is highly relevant in the context of the

discussion of a prolongation of the *Solidarity Pact II* assigning support exclusively to regions located in the East.

Even though the application in this work concentrates on Germany, the theory is easily transferable to the data of other countries attending the HFCS as well as other surveys that use multiple imputation in order to account for item non-response and have a similar data structure. For the imputation, considering the survey design and the explanation of domain differences is important. Furthermore, the approach can also improve the country results for single components of net wealth in the cross-country comparison of the HFCN. For instance, estimates for various financial assets are either not reported for some countries because the sample size is below 25 or are very imprecise (Household Finance and Consumption Network, 2016b).

For further research, it is of interest if the proposed FH approach can also be used for other indicators. The application of the mean enables the usage of Rubin's rule. However, it is unclear if the rule can also be applied for non-linear poverty and inequality indicators like the headcount ratio (Foster et al., 1984) or the Gini coefficient (Gini, 1912). One way could be to use a transformation for indicators that do not fulfill the normality requirement before applying Rubin's rule as supposed in Marshall et al. (2009). A suitable transformation for the headcount ratio might be the arcsin transformation which is also used in the FH approach when the dependent variable is between 0 and 1 (Casas-Cordero et al., 2016; Schmid et al., 2017). In this work, we propose an easy-to-apply approach by using the log-transformation for meeting the model assumptions. Another approach to handle the skewed data could be assuming a skewed normal distribution in the FH model (Moura et al., 2017). Furthermore, future approaches could consider the panel structure of the survey.

Acknowledgments

Kreutzmann and Schmid gratefully acknowledge support by the German Research Foundation within the project QUESSAMI (281573942). Both and also Salvati were further supported by the MIUR-DAAD Joint Mobility Program (57265468). The work of Salvati has also been developed under the support of the Progetto di Ricerca di Ateneo From survey-based to register-based statistics: a paradigm shift using latent variable models' (grant PRA2018-9). This work uses data from the Deutsche Bundesbank Panel on Household Finances. The results published and the related observations and analysis may not correspond to results or analysis of the data producers. We would like to thank the FDSZ and the PHF team at Deutsche Bundesbank for support.

Supplementary material A

Household balance sheet

Table A.1: A household balance sheet.

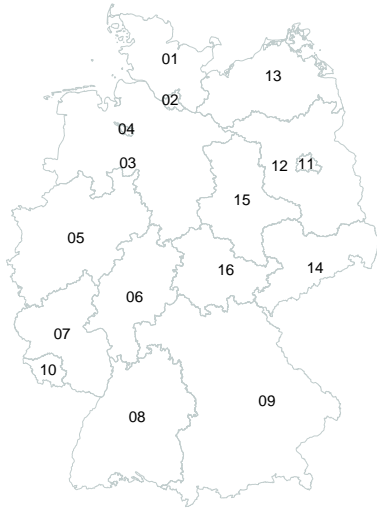
Assets	Liabilities
Real assets	- Outstanding amount of household main residence mortgages and other real estate property mortgages
- Household main residence (HMR)	- Outstanding amount of debt on credit cards and credit lines/bank overdrafts
- Other real estate property	- Outstanding amounts of other, non-collateralized, loans (including loans from commercial providers and private loans)
- Vehicles	
- Valuables	
- Self employment business wealth	
Financial assets	
- Deposits (sight accounts, saving accounts)	
- Mutual funds	
- Bonds	
- Shares (publicly traded)	
- Managed investment accounts	
- Money owed to household	
- Voluntarily private pensions/ Whole life insurance	
- Other financial assets: options, futures, index certificates, precious metals, oil and gas leases, future proceeds from a lawsuit or estate that is being settled, royalties or any other	

Data sources

Table A.2: Variables that potentially help to predict HH net wealth with corresponding references. The numbers in parenthesis state the number of variables for this group.

Variables	Year	Level	Source
Ownershiprate	2011	BL	Landesamt für Statistik Niedersachsen (2014)
Rental and purchase prices per sqm, level of interior (3)	2014	BL ROR	Empirica ag (2017)
Number of houses (2) and types of buildings (6)	2015	ROR	RWI;microm (2016c)
Density of cars and car segment (11)	2015	ROR	RWI;microm (2016b)
Saving quota of HH	2014	BL	Statistische Ämter des Bundes und der Länder (2014d)
Default probability (8)	2015	ROR	RWI;microm (2016e)
Private debtors per 100 inhabitants	2014	ROR	Bundesinstitut für Bau-, Stadt-, und Raumforschung (2017)
Disposable income of private HHs per inhabitant	2014	BL	Statistische Ämter des Bundes und der Länder (2014e)
Average HH income per inhabitant	2014	ROR	Bundesinstitut für Bau-, Stadt-, und Raumforschung (2017)
Unemployment rate,	2014	BL	Statistische Ämter des Bundes und der Länder (2014a)
percentage of employees, civil servants and self-employed	2011	BL	Statistische Ämter des Bundes und der Länder (2011)
Age groups (4)	2014	BL	Statistische Ämter des Bundes und der Länder (2014c)
Age groups (4)	2014	ROR	Bundesinstitut für Bau-, Stadt-, und Raumforschung (2017)
Youth dependency ratio, old-age dependency ratio	2014	BL	Statistische Ämter des Bundes und der Länder (2014b)
Single or couple	2015	ROR	RWI;microm (2016d)

Labels for the regional levels



BL	Name
1	Schleswig-Holstein
2	Hamburg
3	Niedersachsen
4	Bremen
5	Nordrhein-Westfalen
6	Hessen
7	Rheinland-Pfalz
8	Baden-Württemberg
9	Bayern
10	Saarland
11	Berlin
12	Brandenburg
13	Mecklenburg-Vorpommern
14	Sachsen
15	Sachsen-Anhalt
16	Thüringen

Figure A.1: Official municipality keys for the planning regions within its polygon and names of the federal states in Germany.

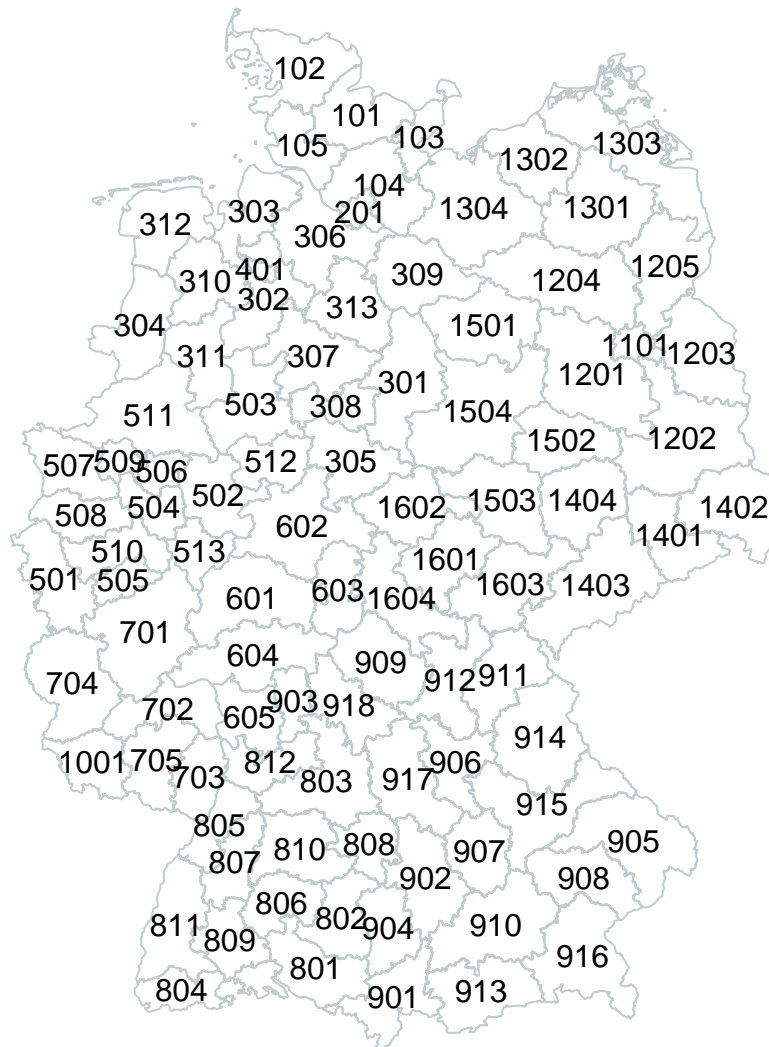


Figure A.2: Official municipality keys for the planning regions within its polygon.

Table A.3: Official municipality keys for the planning regions and corresponding names.

ROR	Name	ROR	Name
101	Schleswig-Holstein Mitte	806	Neckar-Alb
102	Schleswig-Holstein Nord	807	Nordschwarzwald
103	Schleswig-Holstein Ost	808	Ostwürttemberg
104	Schleswig-Holstein Süd	809	Schwarzwald-Baar-Heuberg
105	Schleswig-Holstein Süd-West	810	Stuttgart
201	Hamburg	811	Südlicher Oberrhein
301	Braunschweig	812	Rhein-Neckar (BW)
302	Bremen-Umland	901	Allgäu
303	Bremerhaven	902	Augsburg
304	Emsland	903	Bayerischer Untermain
305	Göttingen	904	Donau-Iller (BY)
306	Hamburg-Umland-Süd	905	Donau-Wald
307	Hannover	906	Industrieregion Mittelfranken
308	Hildesheim	907	Ingolstadt
309	Lüneburg	908	Landshut
310	Oldenburg	909	Main-Rhön
311	Osnabrück	910	München
312	Ost-Friesland	911	Oberfranken-Ost
313	Südheide	912	Oberfranken-West
401	Bremen	913	Oberland
501	Aachen	914	Oberpfalz-Nord
502	Arnsberg	915	Regensburg
503	Bielefeld	916	Südostoberbayern
504	Bochum/Hagen	917	Westmittelfranken
505	Bonn	918	Würzburg
506	Dortmund	1001	Saar
507	Duisburg/Essen	1101	Berlin
508	Düsseldorf	1201	Havelland-Fläming
509	Emscher-Lippe	1202	Lausitz-Spreewald
510	Köln	1203	Oderland-Spree
511	Münster	1204	Prignitz-Oberhavel
512	Paderborn	1205	Uckermark-Barnim
513	Siegen	1301	Mecklenburgische Seenplatte
601	Mittelhessen	1302	Mittleres Mecklenburg/Rostock
602	Nordhessen	1303	Vorpommern
603	Osthessen	1304	Westmecklenburg
604	Rhein-Main	1401	Oberes Elbtal/Osterzgebirge
605	Starkenburg	1402	Oberlausitz-Niederschlesien
701	Mittelrhein-Westerwald	1403	Südsachsen
702	Rheinhessen-Nahe	1404	Westsachsen
703	Rheinpfalz	1501	Altmark
704	Trier	1502	Anhalt-Bitterfeld-Wittenberg
705	Westpfalz	1503	Halle/S.
801	Bodensee-Oberschwaben	1504	Magdeburg
802	Donau-Iller (BW)	1601	Mittelthüringen
803	Heilbronn-Franken	1602	Nordthüringen
804	Hochrhein-Bodensee	1603	Ostthüringen
805	Mittlerer Oberrhein	1604	Südthüringen

Chapter 2

The `fayherriot` command for estimating small-area indicators

2.1 Introduction

Various national and international institutions including the United Nations (Leadership Council of the Sustainable Development Solutions Network, 2015) and the Organisation for Economic Co-operation and Development (OECD) (Piacentini, 2014) collect comprehensive indicator sets for monitoring purposes. Many indicators refer to sub-national areas or domains: federal states, economic sectors, societal groups, etc.

In the socio-economic context, domain-level indicators are usually derived from population surveys by direct estimation. Direct estimates are only based on the survey data and therefore, small sample sizes can limit their precision. For this reason, institutions that provide these kinds of indicators usually require a minimum number of observations per domain or impose limits on the variability of the estimates (eurostat, 2013a; Tzavidis et al., 2018). Furthermore, direct estimates cannot be obtained for out-of-sample domains, i.e., domains without any observation in the sample.

Small area estimation (SAE) techniques use auxiliary data from additional data sources to improve the precision of survey-based direct estimates. One popular approach is the Fay-Herriot model (Fay and Herriot, 1979) due to its moderate data requirements.¹ Direct estimates and auxiliary data are only needed on the domain level.

The command `fayherriot` provides empirical best linear unbiased predictors (EBLUP), a linear combination of the domain-level direct estimator and a regression-synthetic component based on a linear model. The underlying model can also be expressed as a special linear mixed model. In contrast to a standard linear mixed model (encompassed in `mixed` (Rabe-Hesketh and Skrondal, 2012) or `glamm` (StataCorp, 2017)), the Fay-Herriot model builds on two error terms on the same level, the domain level, with domain-specific variances of one error term. The model assumes linearity and normality of its two error terms.

`fayherriot` performs the following:

¹Applications include, e.g., the estimation of income and poverty rates (Powers et al., 2008; Huang and Bell, 2012) and educational indicators (Schmid et al., 2017).

- The Fay-Herriot model as described in Rao and Molina (2015, pp. 123-129) with restricted maximum likelihood and maximum likelihood estimation of the variance of the random effects.
- Estimation of the mean squared error (MSE) as proposed in Datta and Lahiri (2000) and Prasad and Rao (1990).
- Prediction and MSE estimation for out-of-sample domains (Rao and Molina, 2015, p. 126 and p. 139).
- Adjusted estimation methods as proposed in Li and Lahiri (2010) and Yoshimori and Lahiri (2014) to deal with non-positive estimates of the variance of the random effects.
- The log-transformed Fay-Herriot model including a bias-correction by Slud and Maiti (2006) to deal with violations of model assumptions, e.g., non-normality of the error terms.

2.2 The Fay-Herriot model

2.2.1 Modeling

The main idea of the Fay-Herriot (FH) model (Fay and Herriot, 1979) is to combine domain-level direct estimators (based on survey data) with aggregated domain-level covariates (e.g., from register or administrative data). The direct estimator should be a linear statistic such as an arithmetic mean or total.

The FH model builds on a sampling and a linking model. According to the sampling model,

$$\hat{\theta}_d = \theta_d + e_d \quad \text{for } d = 1, \dots, D,$$

the observed direct estimator for domains $d = 1, \dots, D$, $\hat{\theta}_d$, is composed of the true value, θ_d , and a sampling error, e_d , with mean zero and variance $\sigma_{e_d}^2$. The model assumes that the sampling error variance of each domain is known. In practice, the variance of the direct estimator is used frequently as an estimate for $\sigma_{e_d}^2$ (You and Chapman, 2006).

According to the linking model,

$$\theta_d = \mathbf{x}_d^\top \boldsymbol{\beta} + u_d \quad \text{for } d = 1, \dots, D,$$

the true value, θ_d , is explained by domain-specific covariates, \mathbf{x}_d , a random effect, u_d , and regression parameters $\boldsymbol{\beta}$. The random effect is independent identically and normally distributed with mean zero and variance σ_u^2 . The model assumes inter-domain correlations to be zero.

Combining the sampling and the linking model gives the FH model, a linear mixed model of the form,

$$\hat{\theta}_d = \mathbf{x}_d^\top \boldsymbol{\beta} + u_d + e_d \quad \text{for } d = 1, \dots, D. \quad (2.1)$$

The FH estimator (EBLUP) is given by $\hat{\theta}_d^{\text{FH}} = \mathbf{x}_d^\top \hat{\boldsymbol{\beta}} + \hat{u}_d$. It can also be expressed more intuitively as a weighted average of the direct and a regression-synthetic estimator,

$$\hat{\theta}_d^{\text{FH}} = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) \mathbf{x}_d^\top \hat{\boldsymbol{\beta}}. \quad (2.2)$$

The estimator $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{(\hat{\sigma}_u^2 + \hat{\sigma}_{e_d}^2)}$, the so-called ‘‘shrinkage factor’’, weights the direct estimator and the regression-synthetic part. The weight on the direct estimator decreases with the sampling error variance.

For out-of-sample domains, $\hat{\gamma}_d$ is not defined and the regression-synthetic estimator $\mathbf{x}_d^\top \hat{\boldsymbol{\beta}}$ is used.

2.2.2 Estimating the variance of the random error

The FH model requires an estimation of the variance of the random error, σ_u^2 , and of the regression parameters, $\boldsymbol{\beta}$. Standard estimation techniques for σ_u^2 are, among others, restricted maximum likelihood (REML) and maximum likelihood (MLE). These methods do not guarantee positive variance estimates (Li and Lahiri, 2010; Yoshimori and Lahiri, 2014). Especially in case of a small number of domains, the variance estimates can be negatively biased or even below zero. In the latter case, the variance estimate is set to zero. An underestimation of the variance component could lead to a significant over-shrinkage of the direct estimator to the regression-synthetic part, i.e., too much weight is put on the regression-synthetic part. Adjusted estimation methods such as the adjusted maximum residual likelihood approach (ARYL) following Yoshimori and Lahiri (2014) and the adjusted maximum profile likelihood (AMPL) following Li and Lahiri (2010) ensure strictly positive variance estimates.

`fayherriot` allows the estimation of σ_u^2 with the REML (as default), MLE, ARYL and AMPL.² The method can be specified in the command option `sigmamethod`. The vector of regression parameters, $\boldsymbol{\beta}$, is estimated by the empirical best linear unbiased estimator $\hat{\boldsymbol{\beta}}$ (Rao and Molina, 2015, p. 124).

2.2.3 Evaluating the precision

The precision of the EBLUP is evaluated by means of the MSE, defined as:

$$MSE(\hat{\theta}_d^{\text{FH}}) = E \left[\left(\hat{\theta}_d^{\text{FH}} - \theta_d \right)^2 \right].$$

Since the true value θ_d is unobserved, $MSE(\hat{\theta}_d^{\text{FH}})$ must be estimated. For in-sample domains, MSE estimators have been proposed for estimates of σ_u^2 relying on REML (Prasad and Rao, 1990, p. 167), MLE (Datta and Lahiri, 2000, p. 619), ARYL (Yoshimori and Lahiri, 2014), and AMPL (Li and Lahiri, 2010, p. 886). For out-of-sample domains, MSE estimators have been proposed for REML and MLE only (Rao and Molina, 2015, p. 139). `fayherriot` automatically selects the appropriate MSE estimator.

²See Yoshimori and Lahiri (2014) for a general discussion of the comparative advantages of each of the methods.

2.2.4 Dealing with model-assumption violations

The FH model assumes linearity and normality of its two error terms. In case of a violation of these assumptions, a log-transformation of the direct estimator might be an option (Slud and Maiti, 2006). Choosing this option requires an appropriate transformation of the variance of the original direct estimator.³ Neves et al. (2013) suggest the transformation,

$$\begin{aligned}\hat{\theta}_d^* &= \log(\hat{\theta}_d), \\ \text{var}(\hat{\theta}_d^*) &= \left(\hat{\theta}_d\right)^{-2} \text{var}(\hat{\theta}_d),\end{aligned}\tag{2.3}$$

with * indicating the transformed scale.

Equation 2.1 is estimated using $\hat{\theta}_d^*$ as direct estimator and $\text{var}(\hat{\theta}_d^*)$ as estimate for the sampling error variance. To bring the estimated EBLUP and MSE back from the transformed to the original scale, a bias-correction is advised (Slud and Maiti, 2006; Sugawasa and Kubokawa, 2017). `fayherriot` includes two back-transformation methods: the “crude” method, shown in Neves et al. (2013); Rao and Molina (2015), and, as default, the bias-correction proposed by Slud and Maiti (2006). For the point estimates these are defined as follows:

$$\begin{aligned}\hat{\theta}_d^{\text{FH, crude}} &= \exp\left\{\hat{\theta}_d^{\text{FH}*} + 0.5\text{MSE}(\hat{\theta}_d^{\text{FH}*})\right\}, \\ \hat{\theta}_d^{\text{FH, Slud-Maiti}} &= \exp\left\{\hat{\theta}_d^{\text{FH}*} + 0.5\hat{\sigma}_u^2(1 - \hat{\gamma}_d)\right\},\end{aligned}$$

with * indicating the transformed scale.

The Slud-Maiti back-transformation relies on MLE for the estimation of σ_u^2 . Since it requires an estimate of $\hat{\gamma}_d$, it is only applicable for in-sample domains. The “crude” back-transformation can be used for in- and out-of-sample predictions.

For estimating the precision of the back-transformed EBLUPs, Slud and Maiti (2006, p. 248) develop an MSE estimator when using the log-transformation. The “crude” method uses the estimates in the transformed scale and the following back-transformation:

$$\text{MSE}(\hat{\theta}_d^{\text{FH, crude}}) = \exp\left\{\hat{\theta}_d^{\text{FH}*}\right\}^2 \text{MSE}(\hat{\theta}_d^{\text{FH}*})$$

2.2.5 Overview of functionalities

Figure 2.1 gives an overview of the functionalities of the `fayherriot` command.

³It is not appropriate to take the logarithm of the variance. This is because the variance of a log-transformed variable is different from the log-transformed variance of the original variable.

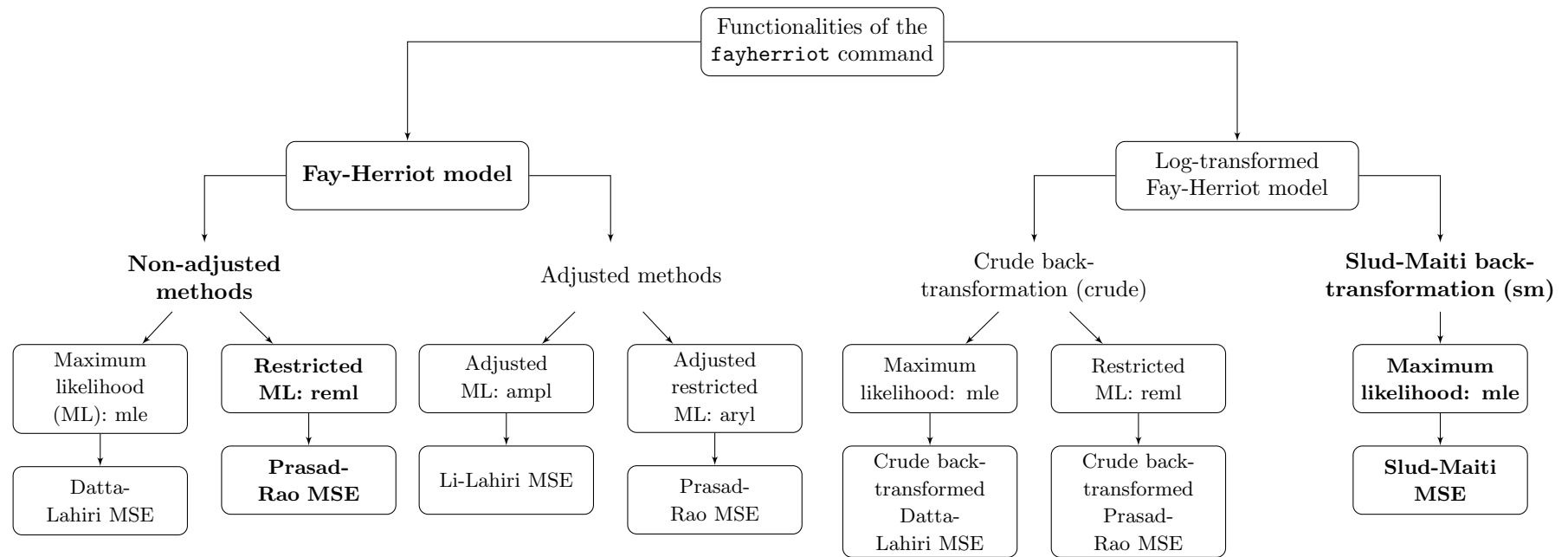


Figure 2.1: Functionalities of the `fayherriot` command. The two lowest levels describe the estimation methods of σ_u^2 and the corresponding MSE estimators, respectively. The default options are written in bold.

2.3 The fayherriot command

2.3.1 Syntax

`fayherriot` runs in Stata 12 and later versions. The syntax is:

```
fayherriot depvar [varlist] [if] [in], variance(varname) [sigmamethod(method)
  logarithm biascorrection(method) initialvalue(#) eblup(name)
  mse(name) gamma nolog]
```

The command runs on datasets on the domain level with one observation per domain. *depvar* is the direct estimator, $\hat{\theta}_d$ (in documentation *theta*), and *varlist* corresponds to the auxiliary explanatory variables, \mathbf{x}_d (in the documentation *X*).

2.3.2 Options for fayherriot

variance(*varname*) determines the variable containing the sampling error variances, *sigma2_e*. This variance is assumed to be known in the model. However, it often needs to be estimated from the data. One possibility is the usage of the estimated variance of the direct estimator *theta* specified in *depvar* for each domain. Whenever the direct estimator needs to be logarithmized with *theta_log* = $\log(\theta)$, the estimated variance can be modified as *sigma2_e_log* = $\sigma^2_e / (\theta^2)$ (Neves et al., 2013).

sigmamethod(*method*) specifies the method for the estimation of the variance of the random effect: *reml*, *mle*, *ampl*, or *aryl*. The default is the *reml* method. Another standard approach is the *mle* approach. However, whenever a zero estimate is received for the variance, which occurs more likely when the number of domains is small, an adjusted maximum likelihood method can help to receive strictly positive values for the variance. Therefore, the *ampl* by Li and Lahiri (2010) and the *aryl* method suggested by Yoshimori and Lahiri (2014) are additionally provided.

logarithm indicates that the dependent variable in *depvar* is the log-transformed direct estimator. A log-transformed Fay-Herriot model is suitable when the linearity or normality assumption of the error terms is not fulfilled. *logarithm* automatically back-transforms EBLUP and MSE to the original scale.

biascorrection(*method*) determines the method for the back-transformation of EBLUP and MSE in a log-transformed Fay-Herriot model. The EBLUPs and MSEs in the transformed scale can be back-transformed using the bias-correction proposed by Slud and Maiti (2006), which is set as a default, and a crude bias-correction (Neves et al., 2013; Rao and Molina, 2015).

initialvalue(#) sets the initial value of the optimization algorithm for estimating the variance of the random effect *sigma2_u* to #. The default value is .0.

eblup(*name*) stores the EBLUP estimates in the variable *name*. For in-sample domains, the EBLUPs are defined as $eblup = X\beta + u$, where *Xbeta* are the estimated fixed effects and *u* is the estimated random effect. The EBLUP can also be expressed as weighted average of the direct estimator and a synthetic part $eblup = \gamma \times \theta + (1 -$

$gamma) Xbeta$. For out-of-sample domains, the EBLUP shrinks to the synthetic part $eblup = Xbeta$.

`mse (name)` stores the MSE estimates in the variable `name`. The MSE depends on the estimation procedure of `sigma2_u`. For `reml`, the MSE estimator relies on Prasad and Rao (1990, p. 167); for `mle` on Datta and Lahiri (2000, p. 619); for `ampl` on Li and Lahiri (2010, p. 886); and for `aryl` method on Yoshimori and Lahiri (2014). For the log-transformed Fay-Herriot model under the Slud-Maiti bias-correction, the MSE is defined as in Slud and Maiti (2006, p. 248). It is only applicable to in-sample domains. Under the crude bias correction, for in- and out-of-sample domains: $mse(eblup_backtransformed) = exp(eblup)^2 \times mse(eblup)$, where `eblup` is in the log scale (Neves et al., 2013).

`gamma` reports summary statistics of the shrinkage factor, $gamma = sigma2_u / (sigma2_u + sigma2_e)$, where `sigma2_u` is the estimated variance of the random effect and `sigma2_e` is the sampling error variance of each domain provided in `variance()`.

`nolog` suppresses the display of the iteration log of the optimization algorithm.

2.3.3 predict after fayherriot:

The syntax for `predict` following `fayherriot` is:

```
predict [type] newvarname [if] [in] [, eblup mse ehat estandard uhat
      gamma cvdirect cvfh]
```

Per default `predict` provides the EBLUPs. Options are:

`eblup` generates the EBLUPs as defined above, the default.

`mse` generates estimates for the MSE as defined above.

`ehat` calculates the residuals. The residuals are defined as $e = (1 - gamma) \times (theta - Xbeta)$, where `theta` is the direct estimator given in `depvar`.

`estandard` calculates the standardized residuals defined as $e / sqrt(sigma2_e)$, where `sigma2_e` is the sampling error variance in `variance(varname)`.

`uhat` calculates the random effects. The random effects are defined as: $u = gamma \times (theta - Xbeta)$.

`gamma` generates the shrinkage factor as defined above.

`cvdirect` calculates the coefficient of variation of direct estimates. $cvdirect = 100 \times sqrt(sigma2_e) / theta$, where `theta` corresponds to `depvar` and `sigma2_e` is the sampling error variance provided in `variance(varname)`. In case `logarithm` is specified, $cvdirect = 100 \times sqrt(sigma2_e') / theta'$ with $theta' = exp(theta_log)$, and $sigma2_e' = var(theta_log) \times (theta')^2$.

`cvfh` calculates the coefficient of variation based on EBLUPs: $cvfh = 100 \times sqrt(mse) / eblup$.

2.3.4 Stored results

Scalars

<code>e(N_in)</code>	number of observations used for estimation of <code>e(b)</code> and <code>e(sigma2_u)</code>	<code>e(N_out)</code>	number of out-of-sample observations for which EBLUP is calculated
<code>e(sigma2_u)</code>	estimated <code>sigma2_u</code>	<code>e(r2_a)</code>	adjusted R-squared
<code>e(r2_fh)</code>	adjusted R-squared according to Lahiri and Suntornchost (2015)	<code>e(p_e)</code>	p-value of Shapiro-Wilk test for normality of residuals
<code>e(V_e)</code>	index of Shapiro-Wilk test statistic, test for normality of residuals	<code>e(p_u)</code>	p-value of Shapiro-Wilk test statistic, test for normality of <code>u</code>
<code>e(V_u)</code>	index of Shapiro-Wilk test for normality of <code>u</code>		

Macros

<code>e(cmd)</code>	<code>fayherriot</code>	<code>e(title)</code>	Fay-Herriot estimation
<code>e(depvar)</code>	name of dependent variable	<code>e(variance)</code>	name of variance variable
<code>e(sigma_method)</code>	<code>sigmau_2</code> estimation method	<code>e(bias_correction)</code>	bias correction method for the back-transformation of transformed EBLUPs
<code>e(logarithm)</code>	Logarithm true or false		
<code>e(predict)</code>	program to implement predict	<code>e(properties#(b V))</code>	

Matrices

<code>e(b)</code>	coefficient vector	<code>e(V)</code>	variance-covariance matrix of coefficients
<code>e(gamma)</code>	summary of values of shrinkage factor <code>gamma</code>		

Functions

<code>e(sample)</code>	marks estimation sample
------------------------	-------------------------

2.4 Example

We use the FH model to estimate households' material well-being in 2015 in Germany: at the level of federal states (16 divisions), planning regions (96 divisions), and districts (402 divisions). Material well-being is defined as region-specific average equivalent income, i.e., household disposable income divided by the OECD modified scale (Hagenaars et al., 1994).

Following the policies used by several statistical agencies to evaluate the precision of the regional estimates, we rely on the coefficient of variation (CV), the standard error of the estimate divided by the estimate (in percent). For instance, Statistics Canada releases data without warning about low precision if the CV is below 16.5 percent (Statistics Canada, 2013; eurostat, 2013a).

2.4.1 Data description and direct estimates

We derive the direct estimates from the German Socio-Economic Panel (SOEP), a household survey covering about 15,000 households per year (Goebel et al., 2018).

Table 2.1 provides the division-specific numbers of SOEP households. Sample sizes by federal states are large (median: 624), ranging from 114 to 3,159 observations. Sample sizes by planning regions are considerably smaller (median: 132), ranging from 32 to 665 observations. Sample sizes by districts range from 10 to 648 observations (median: 32).⁴ Due to small sample

⁴For confidentiality issues, we discarded areas with fewer than 10 observations. This left us with 357 out of 402 districts.

Table 2.1: Number of regions and sample sizes.

Regional division	Number of regions	Sample size distribution				
		Min	p10	p25	p50	Max
Federal states	16	114	144	444	624	3158
Planning regions	96	32	61	88	132	665
Districts	357	10	14	20	32	648

Note: Data are from SOEP v33.1. Own computations.

Table 2.2: Summary of mean equivalent household income and coefficients of variation by regional level.

Regional division	Min	p10	p25	p50	p75	p90	Max
(A) Mean equivalized household income							
Federal states	1362	1398	1492	1683	1777	1841	1863
Planning regions	1298	1400	1495	1664	1780	1898	2101
Districts	1023	1311	1463	1641	1847	2049	2976
(B) Coefficient of variation							
Federal states	0.6	0.8	1.4	2.2	3.8	6.4	8.0
Planning regions	1.5	3.4	4.1	5.3	7.2	9.0	18.2
Districts	2.2	5.8	7.6	10.2	13.6	16.7	42.5

Note: Data are from SOEP v33.1. Own computations.

sizes, we expect that many direct estimates for planning regions and districts are measured with high imprecision.

For each regional level, Table 2.2 provides direct estimates of mean equivalent income and coefficients of variation, our precision indicator.⁵ The table suggests considerable regional heterogeneities in material well-being. Across federal states, mean equivalent income ranges from €1,362 to €1,863; across planning regions from €1,298 to €2,101; and across districts from €1,023 to €2,976. As expected, coefficients of variation increase as we move to smaller regional levels. In line with the policy of Statistics Canada, not all estimates could be reported for the planning regions and the districts without warning of low precision. In the following, we show how this can be achieved using the FH model. In particular, a) the precision of all estimates will be improved, and b) estimates for the districts without direct estimator can be received.

2.4.2 Estimation using fayherriot

For estimating the FH model, we rely on the direct estimates of average equivalent incomes (Table 2.2), their sampling error variances, $\sigma_{e_d}^2$, and region-specific explanatory variables. The set of explanatory variables in this example includes the unemployment rate, the share of population older than 65 years, and per-capita income tax revenue.⁶

⁵We estimated standard errors using the random group estimator to account for the survey sampling design (Rendtel, 1995).

⁶The explanatory variables are obtained from INKAR (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017), a database of regional indicators derived from high-quality and large-scale national census and register data.

FH model for the planning regions

In the following, we detail the application of `fayherriot` at the level of planning regions. In this example, all regions are sampled and the model assumptions are fulfilled. The underlying dataset includes 96 observations, one observation per region:

```
. use dataROR.dta, clear
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
income	96	1658.387	188.7142	1297.915	2100.683
directvari_e	96	11448.52	12856.35	612.4922	96107
unemployment	96	6.259375	2.579212	2.1	12.8
incometax	96	399.2719	105.6913	211.6	705
share65	96	56.48438	.8259385	54.9	58.2
N	96	162.3854	125.7412	32	665

To estimate the FH model, we type:

```
. fayherriot income unemployment incometax share65, ///
> variance(directvariance) gamma nolog
```

Sigma2_u estimation method: reml	N in sample =	96
Log dependent variable: No	N out of sample =	0
EBLUP and MSE bias correction: None	Sigma2_u =	4683.7186
	Adj R-squared =	0.5769
	FH R-squared =	0.7808

Gamma				
Min	5%	Median	95%	Max
0.0465	0.1464	0.3726	0.7307	0.8844

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
income					
unemployment	5.956308	6.664691	0.89	0.371	-7.106247 19.01886
incometax	1.278903	.1365014	9.37	0.000	1.011365 1.546441
share65	-38.88106	18.04844	-2.15	0.031	-74.25536 -3.506763
_cons	3301.427	1013.564	3.26	0.001	1314.878 5287.976


```
Shapiro-Wilk test for normality:
Residuals e (standardized) V = 0.837 p-value = 0.653
Random effects u V = 0.392 p-value = 0.981
```

The syntax of the command is in line with the familiar Stata regression syntax: `income` contains the direct estimates of mean equivalent income and is regressed on the regional explanatory variables, `unemployment`, `incometax`, and `share65`. The `variance` option specifies the variable containing the sampling error variances, `directvariance`. We specify the `gamma` option to display summary statistics of shrinkage factors $\hat{\gamma}_d$. `nolog` suppresses the iteration log of the optimization algorithm.

`N in sample` indicates that the full set of 96 planning regions was used in the estimation. `FH R-squared` is an indicator for the goodness of fit of the FH model, proposed by Lahiri and Suntornchost (2015, p. 317, $Adj R_h^2$). Similar to the standard R^2 , it expresses the explained variation of `income` in relation to the total variation, while taking into account that some variation in `income` is due to the sampling error. In this example, about 78% of the variation

is explained.

The variance of the random effects, $\hat{\sigma}_u^2 = 4,683.7186$, is estimated using the REML approach (default). Together with the sampling error variances $\sigma_{e_d}^2$, it determines the shrinkage factor $\hat{\gamma}_d$. The shrinkage factor shows how direct estimates and model predictions are weighted when calculating the EBLUP. Large values of $\hat{\gamma}_d$ mean that a large weight is given to the direct estimate $\hat{\theta}_d$. In our example, the distribution of $\hat{\gamma}_d$ ranges from 0.0465 to 0.8844 with its median being 0.3726. So for some regions, the EBLUP relies strongly on the model predictions (small value of $\hat{\gamma}_d$), and strongly on the direct estimator for others (large value of $\hat{\gamma}_d$). The Shapiro-Wilk test for normality shows that neither normality of the realized residuals, \hat{e}_d , nor of the random effects, \hat{u}_d , is rejected. Hence, the model assumptions are not violated.

Log-transformed FH model for the districts

In the district-level analysis, not all regions are sampled and the normality assumption of the model is violated. Hence, we log-transform equivalent incomes and the variances of the sampling error,

```
. use dataDistricts.dta, clear
. gen logincome = log(income)
(45 missing values generated)
. gen directlogvariance = directvariance/income^2
(45 missing values generated)
```

and estimate the log-transformed FH model:

```
. fayherriot logincome unemployment incometax share65, ///
> variance(directlogvariance) nolog logarithm

Sigma2_u estimation method: mle           N in sample =           357
Log dependent variable: Yes              N out of sample =          45
EBLUP and MSE bias correction: sm        Sigma2_u =                0.0089
                                          Adj R-squared =           0.2891
                                          FH R-squared =            0.4745
```

logincome	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
unemployment	-.0004102	.003304	-0.12	0.901	-.0068858	.0060655
incometax	.0007471	.0000904	8.26	0.000	.0005698	.0009243
share65	-.0063528	.003548	-1.79	0.073	-.0133067	.0006011
_cons	7.241288	.1051244	68.88	0.000	7.035248	7.447328

```
Shapiro-Wilk test for normality:
Residuals e (standardized)  V =    1.614  p-value = 0.128
Random effects u            V =    0.830  p-value = 0.670
```

By specifying the `logarithm` option, `fayherriot` transforms the estimated EBLUP and MSE back to the original scale. Because we did not specify the bias-correction method, the estimation method is MLE and the bias correction follows Slud and Maiti (2006) (see Figure 2.1). In this default setting, only estimates for the 357 in-sample districts are calculated. `biascorrection` (*crude*) could be specified to obtain in- and out-of-sample estimates.

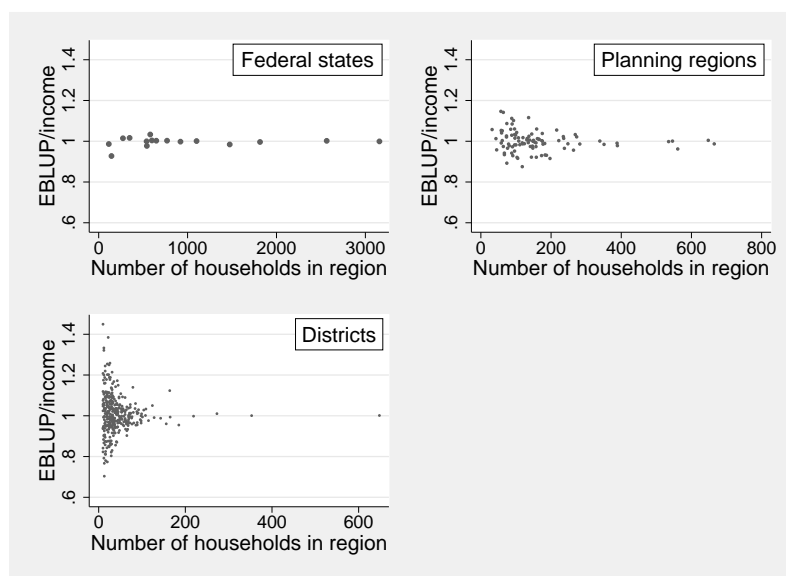


Figure 2.2: Ratio of the EBLUP to the direct estimates plotted against regional sample sizes for all three regional divisions – federal states, planning regions, and districts. Only in-sample domains are plotted. Data are from SOEP v33.1. Own computations.

2.4.3 Comparison of direct and FH estimates

Next we compare the direct with the FH point estimates (EBLUP) and assess their precision. There are two equivalent ways to obtain the EBLUPs and their level of precision (MSE). First, by specifying the `eblup` (*varname*) and `mse` (*varname*) option (here done for the planning regions):

```
. fayherriot income unemployment incometax share65, ///
variance(directvariance) nolog eblup(eblupROR) mse(mseROR)
```

Second, by using the post-estimation `predict` routine directly after the `fayherriot` command:

```
. predict eblupROR, eblup
. predict mseROR, mse
```

An additional feature of `predict` is that it provides the *CV* for the direct and FH estimates.

```
. predict cvROR_FH, cvfh
. predict cvROR_direct, cvdirect
```

To assess the magnitude of adjustments, Figure 2.2 presents the ratios of EBLUPs and direct estimates against region-specific sample sizes.⁷ For federal states, the ratios are all close to one, suggesting small adjustments of the direct estimator. For planning regions and districts, adjustments are larger, an expected result given smaller sample sizes of these domains.

To assess the gain in precision, Figure 2.3 provides boxplots of coefficients of variation for the direct and FH estimates. The horizontal line indicates the threshold of 16.5 suggested by Statistics Canada. For the direct estimates, several CVs at the district and planning region level exceed the threshold. For the FH estimates, in contrast, CVs for all regional levels are under the threshold.

⁷For further comparison methods, we refer to Brown et al. (2001).

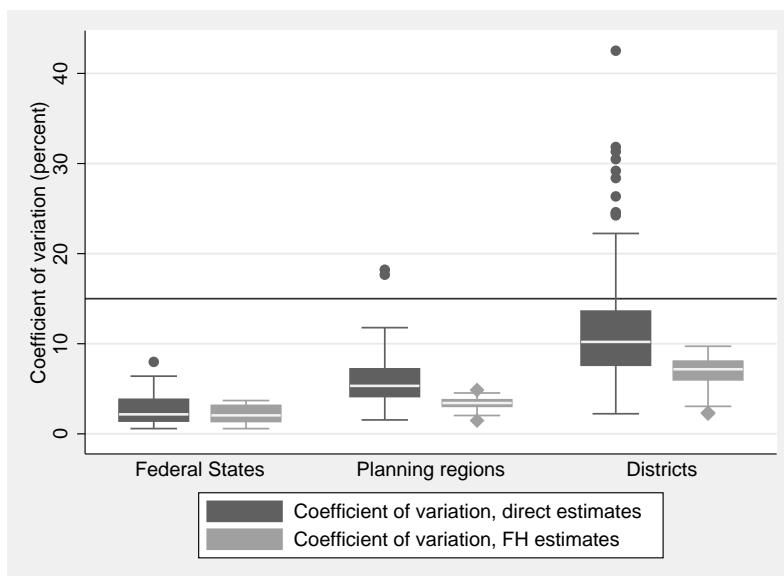


Figure 2.3: Boxplots of the distribution of the coefficients of variation for the federal states, the planning regions, and the districts. The horizontal line indicates the precision threshold of 16.5 percent. Only in-sample domains are plotted. Data are from SOEP v33.1. Own computations.

2.5 Conclusion

SAE techniques are designed to improve the precision of domain indicators. One such technique is the Fay-Herriot (FH) model. It aims at improving the precision of direct estimators from a survey by using additional domain-level covariate information. We introduce the `fayherriot` command and provide an application to regional heterogeneities in material well-being in Germany. The results show that the precision of the FH model estimates is markedly higher than that of the direct estimates.

2.6 Acknowledgments

Halbmeier and Schröder thank Johannes König for his valuable remarks and comments as well as Paul Brockmann, Deborah Anne Bowen, and Fabian Nemeček for excellent research assistance. Kreutzmann and Schmid gratefully acknowledge support by the German Research Foundation within the project QUESSAMI (281573942) and the MIUR-DAAD Joint Mobility Program (57265468).

Part II

Estimation of Disaggregated Non-Linear Indicators based on Income and Wealth Data

Chapter 3

Estimation of sample quantiles: Challenges and issues in the context of income and wealth distributions

3.1 Motivation

The four most popular statistical software packages in terms of the number of scholarly articles in 2016 are SPSS (IBM Corp, 2013), R (R Core Team, 2018), SASTM software (SAS Institute Inc., 2018), and Stata (StataCorp, 2015) (Muenchen, 2017). All of them enable the estimation of sample quantiles which are a well-known statistical measure that helps to describe distributions in many research fields. Generally, users would assume the same result for the quantiles regardless of the statistical software. Table 3.1 shows exemplary results of the median and the quartiles for household net wealth from the Panel of Household Finances (PHF) (Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank, 2014) and for disposable income aggregated at the federal state level, obtained from the national accounts for the federal states (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017).

Table 3.1: Unweighted median and quartiles of household net wealth (left) and disposable income (right). The measures of household net wealth of households in the German states Nordrhein-Westfalen, Rheinland-Pfalz and Saarland are estimated from household data based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations. The measures of disposable income are estimated on aggregated data for the federal states from the database Indikatoren und Karten zur Raum- und Stadtentwicklung (INKAR) (Bundesinstitut für Bau-, Stadt-, und Raumforschung, 2017).

	Household net wealth (n = 1144)			Disposable income (n = 16)		
	0.25	0.5	0.75	0.25	0.5	0.75
R	21275	172780	409383	1532	1672	1798
SAS	20950	172780	409856	1525	1672	1802
SPSS	20300	172780	408910	1512	1672	1793
Stata	20950	172780	409856	1525	1672	1802

It is noticeable that the results differ across software programs even when the sample size is relatively large with 1,144 observations. This is due to the fact that the software programs use different quantile definitions. Attempts to encourage a common quantile definition have so far not been successful (Hyndman and Fan, 1996; Langford, 2006). Therefore, this work describes and compares the different quantile definitions that are implemented in the aforementioned statistical software programs.

In theory, quantiles are defined by the inverse of the cumulative distribution function (cdf), also called the quantile function, if the distribution function is known and invertible. The quantile function Q can be expressed as

$$Q(p) = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\} \quad \text{for } p \in (0, 1),$$

where $F(x)$ is the distribution function and p the quantile level. It returns the minimum value of x from amongst all those values whose cdf value is at least equal to p . In case the theoretical distribution is not known and data is finite, e.g., when the data source is a survey, quantiles need to be estimated. We denote X_1, \dots, X_n as n independent and identically distributed sample observations where $X_{(1)}, \dots, X_{(n)}$ are the corresponding order statistics, i.e., the sample observations in ascending order (David and Nagaraja, 2003). Based on these order statistics, most quantile definitions can be classified into two categories:

1. Sample quantiles that are defined as order statistics or weighted averages of two order statistics:

$$Q_p = (1 - \gamma)X_{(i)} + \gamma X_{(i+1)},$$

where $X_{(i)}$ is the i th order statistic and the value of γ is a weighting factor (often a function of i).

2. Sample quantiles that make use of a weighted average of all order statistics with different weighting factors:

$$Q_p = \sum_{i=1}^n W_i X_{(i)},$$

where $X_{(i)}$ is the i th order statistic and W_i a weighting factor depending on i .

The estimation of the category 1 type of quantiles goes back to at least the 19th century (see e.g., Galton (1889); Edgeworth (1886) and for a more detailed review see Eubank (2004)). Many different definitions have been developed since then. They basically differ in the way in which the two order statistics are weighted, i.e., how γ is defined (Hazen, 1914; Weibull, 1939; Gumbel, 1939; Parzen, 1979). Among others, Cramér (1946) and Chatterjee (2011) study the asymptotic properties of sample quantiles and show the asymptotic normality for the inverse of the empirical cumulative distribution function (ecdf). Along with advances in computation, quantile estimators of category 2 were introduced (Harrell and Davis, 1982; Yang, 1985; Sheather and Marron, 1990; Fan et al., 2014), especially in order to improve the efficiency of

quantile estimators regardless of the underlying distribution. Recently, Sfakianakis and Verginis (2008) introduced a new family of quantile estimators and Makkonen and Pajari (2014) proposed the usage of the true rank probabilities for defining sample quantiles. The definitions mentioned above are nonparametric, i.e., no distributional assumption about the population quantile is made. For semi-parametric and parametric approaches we refer, for instance, to Hosking (1990), Longford (2011) and Wei et al. (2015). This review of quantile definitions emphasizes the wide range of definitions. However, by far, not all known definitions are available in standard statistical software.

This work focuses on the definitions that are implemented in the programs SPSS, R, SASTM software, and Stata. Its purpose is

- to draw attention to the different quantile definitions in statistical software,
- to discuss the possibilities for the inclusion of sampling weights and variance estimation,
- to compare the performance of different quantile estimators using a simulation study based on a theoretical distribution that approximates the German income distribution,
- and to discuss the challenges when evaluating the sample quantiles of a skewed empirical distribution of net wealth.

The first and the third point extend the work of Parrish (1990), Dielmann et al. (1994) and Hyndman and Fan (1996). While Parrish (1990) investigates a range of different quantile definitions available in software when the distribution is normal, Dielmann et al. (1994) also consider symmetric long-tailed and skewed distributions. Both suggest a quantile estimator proposed by Harrell and Davis (1982) for symmetric distributions and Parrish (1990) further states that a quantile definition by Hazen (1914) is least biased. Hyndman and Fan (1996) advise standardizing the quantile definition in statistical software. Their decision is based on different theoretical properties of quantiles.

Since many social indicators are based on income and wealth data (Beste et al., 2018; eurostat, 2018b), these distributions are used for the evaluation of the different quantile definitions in this work. In line with the literature (see e.g., Bhat, 1994; Marchetti et al., 2018), income and wealth distributions are assumed to be continuous in this work. For quantile estimators of discrete random variables, its properties and implementation in R, see Ma et al. (2011) and Geraci (2016). In particular, the interest in the disaggregated information of these indicators has increased in recent decades, i.e., indicators are estimated for regional, sociodemographic or other subgroups of the population. Even if the sample size is large for the whole population, the sample size might be small for the subgroup of interest. Therefore, the quantile estimators are compared in relatively small samples, which is in line with Dielmann et al. (1994). Since income and wealth data is often provided by surveys, we also discuss the possibilities to incorporate sampling weights in the different software programs. This aspect is not covered in any of the other studies. As a theoretical development, the quantile definition by Harrell and Davis (1982), that shows convincing results in Parrish (1990) and Dielmann et al. (1994), is extended in this work in order to account for sampling weights. A short presentation of variance estimation for quantiles complements this overview of quantile estimation in statistical software.

The paper is structured as follows. Section 3.2 introduces the default options of the four software programs and additional definitions recommended by the literature. Furthermore, the feasibility of incorporating sampling weights and methods to produce variance estimates are discussed. In Section 3.3, two simulation studies are conducted for the comparison of the quantile definitions. The first uses a theoretical distribution that approximates income data and the second is a design-based simulation based on data from the PHF (Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank, 2014). Section 3.4 summarizes the results.

3.2 Quantile definitions in statistical software

In this section, the quantile definitions evaluated in the simulation studies are introduced. As already mentioned, the default quantile definitions, i.e., the quantile definitions that are chosen when the user does not change it explicitly, in the software programs SPSS, R, SASTM software and Stata are investigated. Particularly, the focus lies on the function EXAMINE in SPSS, the `quantile` function in R, the procedure UNIVARIATE in SASTM software, and the `pctile` command in Stata (see code examples in the online supplementary material of this paper). These functions allow the use of five, nine, five, and two different quantile definitions, respectively. SPSS allows the quantile definition to be altered in the function EXAMINE by adding the name of the definition. In R, the quantile definition can be changed by using the argument `type` in the `quantile` function. SASTM software allows changes to the definition under the option PCTLDEF. Similarly, the alternative option `altdef` can be added to the `pctile` command in Stata. All quantile definitions implemented in the mentioned functions are stated in Table 3.2. In the following, the default options of these functions and definitions suggested by the existing literature are presented.

3.2.1 The default options and suggestions from the literature

Since the theoretical quantile is the inverse of the cdf, an intuitive definition of a sample quantile is the inverse of the empirical cumulative distribution function (ecdf). Within the definitions based on the ecdf, the **inverse of the ecdf with averaging at discontinuities (Q1)** is probably the most popular quantile definition and is defined as

$$Q1_p = \begin{cases} X_{(1)} & \text{if } p = 0; \\ 0.5 (X_{(i)} + X_{(i+1)}) & \text{if } 0 < p < 1 \quad \text{and} \quad g = 0; \\ X_{(i+1)} & \text{if } 0 < p < 1 \quad \text{and} \quad g \neq 0; \\ X_{(n)} & \text{if } p = 1, \end{cases}$$

where $i = \lfloor np \rfloor$ and $g = np - i$. The floor function $\lfloor \cdot \rfloor$ returns the greatest integer less than or equal to its argument.

Among others, Cramér (1946) shows the asymptotic normality of the $Q1$ definition when np is not an integer. The quantile definition $Q1$ is the default option of command `pctile` in Stata and of the UNIVARIATE procedure in SASTM software but it is also implemented in the

Table 3.2: Quantile definitions in the EXAMINE function in SPSS, the quantile function in R, the UNIVARIATE procedure in SASTM software and command pctile in Stata. The defaults are written in bold. Abbreviations are only added when the quantile definition is explicitly described in the paper. All other definitions can be found in Appendix B.1.

Description and abbreviation	SPSS EXAMINE	R quantile	SAS UNIVARIATE	Stata pctile
Inverse of the ECDF	Empirical	Type 1	3	
Inverse of the ECDF with averaging at discontinuities	<i>Q1</i> Aempirical	Type 2	5	default
Closest to np	Round	Type 3	2	
Linear interpolation of the ECDF	Waverage	Type 4	1	
Piecewise linear function where the knots are the values midway through the steps of the empirical distribution	<i>Q4</i>	Type 5		
Linear interpolation of the expectations for the order statistics for the uniform distribution on [0,1]	<i>Q2</i> Haverage	Type 6	4	altdef
Linear interpolation of the modes for the order statistics for the uniform distribution on [0,1]	<i>Q3</i>	Type 7		
Linear interpolation of the approximate medians for order statistics	<i>Q5</i>	Type 8		
Approximation to $F(E(X_k))$ for the normal distribution		Type 9		

other software programs. While this quantile function is a discontinuous function, the default options in R and SPSS are continuous functions. The latter uses a **linear interpolation of the expectations for the order statistics for the uniform distribution on [0,1] (Q2)**. In other words, the vertices split the sample into $n + 1$ parts with probability $1/(n + 1)$ on average (Weibull, 1939). The definition can be expressed by

$$Q2_p = \begin{cases} X_{(1)} & \text{if } p < \frac{1}{n+1}; \\ (1 - \gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{1}{n+1} \leq p < \frac{n}{n+1}; \\ X_{(n)} & \text{if } p \geq \frac{n}{n+1}, \end{cases}$$

where $i = \lfloor np_k + p \rfloor$, $p_k = \frac{(np+p)}{n+1}$, $\gamma = np_k + p - i$.

Instead of using linear interpolation of expectations, the default option in R defines quantiles by the **linear interpolation of the modes for the order statistics for the uniform distribution on [0,1] (Q3)**. Hyndman and Fan (1996) emphasize that a desired property of this definition is the division of the sample range into $n - 1$ intervals of which $p100\%$ are on the left of $Q3_p$ and $(1 - p)100\%$ are on the right. It is defined as follows,

$$Q3_p = \begin{cases} (1 - \gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } 0 \leq p < 1; \\ X_{(n)} & \text{if } p = 1, \end{cases}$$

where $i = \lfloor np_k + 1 - p \rfloor$, $p_k = \frac{(np+1-p)-1}{n-1}$, $\gamma = np_k + 1 - p - i$.

In addition to these default options, three more definitions that are available, e.g., in **R**, are included in the comparison due to their promising results in existing studies. Since the most popular statistical software programs use different quantile definitions as a default, Hyndman and Fan (1996) suggest the usage of one common definition. They define six properties that a quantile should fulfill (see Appendix B.1.5) and investigate the nine different definitions stated in Table 3.2 with regard to these properties. Only the **piecewise linear function where the knots are the values midway through the steps of the empirical distribution (Q4)** satisfies all defined properties. Furthermore, Parrish (1990) also determines this quantile definition as least biased, especially for small samples when the underlying distribution is normal. The Q_4 definition that goes back to Hazen (1914) is one option of the `quantile` function in **R** and is defined as

$$Q_{4p} = \begin{cases} X_{(1)} & \text{if } p < \frac{0.5}{n}; \\ (1 - \gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{0.5}{n} \leq p < \frac{n-0.5}{n}; \\ X_{(n)} & \text{if } p \geq \frac{n-0.5}{n}, \end{cases}$$

where $i = \lfloor np_k + 0.5 \rfloor$, $p_k = \frac{(np)}{n}$, $\gamma = np_k + 0.5 - i$.

Nevertheless, Hyndman and Fan (1996) propose the **linear interpolation of the approximate medians for order statistics (Q5)** (Johnson and Kotz, 1970) as a common standard even though this definition only fulfills five of the six properties. The definition is suggested since Q_5 is approximately median-unbiased regardless of the underlying distribution. The definition is also only available in **R** and can be expressed by

$$Q_{5p} = \begin{cases} X_{(1)} & \text{if } p < \frac{2/3}{n+1/3}; \\ (1 - \gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{2/3}{n+1/3} \leq p < \frac{n-1/3}{n+1/3}; \\ X_{(n)} & \text{if } p \geq \frac{n-1/3}{n+1/3}, \end{cases}$$

where $i = \lfloor np_k + \frac{(p+1)}{3} \rfloor$, $p_k = \frac{(np + \frac{(p+1)}{3}) - \frac{1}{3}}{n + \frac{1}{3}}$, $\gamma = np_k + \frac{(p+1)}{3} - i$.

All of these definitions belong to the first category of quantile definitions (see also Table 3.3). However, one definition of category 2, the Harrell-Davis estimator (Harrell and Davis, 1982), is often mentioned in the literature as one of the most efficient estimators (Parrish, 1990; Dielmann et al., 1994; Vélez and Correa, 2014). Therefore, the **Harrell-Davis estimator (Q6)** which is implemented in function `hdquantile` in the **R** package **Hmisc** (Harrell Jr et al., 2018) is also considered in this work. The estimator can be expressed as follows,

$$Q_{6p} = \sum_{i=1}^n W_{n,i} X_{(i)},$$

with weighting factors

$$\begin{aligned} W_{n,i} &= \frac{1}{\beta\{(n+1)p, (n+1)(1-p)\}} \int_{(i-1)/n}^{i/n} y^{(n+1)p-1} (1-y)^{(n+1)(1-p)-1} dy, \\ &= B_{i/n}\{p(n+1), (1-p)(n+1)\} - B_{(i-1)/n}\{p(n+1), (1-p)(n+1)\}, \end{aligned}$$

Table 3.3: Quantile definitions presented in Section 3.2.1 that belong to category 1.

Abbreviation and definition	i	p_k	g/γ
$Q1_p \begin{cases} X_{(1)} & \text{if } p = 0; \\ 0.5(X_{(i)} + X_{(i+1)}) & \text{if } 0 < p < 1, \quad g = 0; \\ X_{(i+1)} & \text{if } 0 < p < 1, \quad g \neq 0; \\ X_{(n)} & \text{if } p = 1; \end{cases}$	$[np]$		$np - i$
$Q2_p \begin{cases} X_{(1)} & \text{if } p < \frac{1}{n+1}; \\ (1-\gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{1}{n+1} \leq p < \frac{n}{n+1}; \\ X_{(n)} & \text{if } p \geq \frac{n}{n+1}; \end{cases}$	$[np_k + p]$	$\frac{(np+p)}{n+1}$	$np_k + p - i$
$Q3_p \begin{cases} (1-\gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } 0 \leq p < 1; \\ X_{(n)} & \text{if } p = 1; \end{cases}$	$[np_k + 1 - p]$	$\frac{(np+1-p)-1}{n-1}$	$np_k + 1 - p - i$
$Q4_p \begin{cases} X_{(1)} & \text{if } p < \frac{0.5}{n}; \\ (1-\gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{0.5}{n} \leq p < \frac{n-0.5}{n}; \\ X_{(n)} & \text{if } p \geq \frac{n-0.5}{n}; \end{cases}$	$[np_k + 0.5]$	$\frac{(np)}{n}$	$np_k + 0.5 - i$
$Q5_p \begin{cases} X_{(1)} & \text{if } p < \frac{2/3}{n+1/3}; \\ (1-\gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{2/3}{n+1/3} \leq p < \frac{n-1/3}{n+1/3}; \\ X_{(n)} & \text{if } p \geq \frac{n-1/3}{n+1/3}; \end{cases}$	$[np_k + \frac{(p+1)}{3}]$	$\frac{(np + \frac{(p+1)}{3}) - \frac{1}{3}}{n + \frac{1}{3}}$	$np_k + \frac{(p+1)}{3} - i$

where $y = F(x)$ and $B_x(a, b)$ denotes the incomplete beta function (Majumder and Bhat-tacharjee, 1973; Phien, 1990). While Harrell and Davis (1982) state that this quantile definition is, under some assumptions, asymptotically normally distributed for all quantile levels, Yoshizawa et al. (1985) argue that this is only true for the median.

3.2.2 Incorporation of sampling weights

Most studies that compare quantile definitions do not discuss the incorporation of sampling weights. However, this is especially important for the representative analysis of surveys since most surveys are based on a complex sampling design and use weighting procedures to adjust for unit non-response and other irregularities (Lohr, 2010; Lavallée and Beaumont, 2015; Steinhauer et al., 2015). In the following, not all quantile definitions described above are also shown with sampling weights. This is due to the fact that the inclusion of sampling weights is not well documented for each definition in the software programs. The function EXAMINE in SPSS has the option to include weights for all quantile definitions provided. For the default, Haverage, the weighted quantile can be expressed by

$$WQ1_p = \begin{cases} X_{(i+1)} & \text{if } g_1^* \geq 1; \\ (1 - g_1^*)X_{(i)} + g_1^*X_{(i+1)} & \text{if } g_1^* < 1 \quad \text{and} \quad w_{(i+1)} \geq 1; \\ (1 - g_1)X_{(i)} + g_1X_{(i+1)} & \text{if } g_1^* < 1 \quad \text{and} \quad w_{(i+1)} < 1, \end{cases}$$

where $X_{(i)}$ is the i th order statistic of sample observations $X = (X_1, \dots, X_n)$ and $w = (w_1, \dots, w_n)$ are the corresponding sampling weights, i satisfying $\sum_{j=1}^i w_j \leq p \left(1 + \sum_{j=1}^n w_j\right) < \sum_{j=1}^{i+1} w_j$, $g_1^* = p \left(1 + \sum_{j=1}^n w_j\right) - \sum_{j=1}^i w_j$ and $g_1 = \frac{g_1^*}{\sum_{j=1}^{i+1} w_j}$.

In Stata, the default of the command `pctile` is defined with weights but the alternative definition (`altdef`) is not. Similarly, the procedure UNIVARIATE only allows for the in-

clusion of weights in its default definition. The weighted quantile option used in the SASTM software and Stata is defined as follows,

$$WQ2_p = \begin{cases} X_{(1)} & \text{if } w_1 < p \sum_{j=1}^n w_j; \\ \frac{1}{2} (X_{(i)} + X_{(i+1)}) & \text{if } \sum_{j=1}^i w_j = p \sum_{j=1}^n w_j; \\ X_{(i+1)} & \text{if } \sum_{j=1}^i w_j < p \sum_{j=1}^n w_j < \sum_{j=1}^{i+1} w_j, \end{cases}$$

where $X_{(i)}$ is the i th order statistic of sample observations $X = (X_1, \dots, X_n)$ and $w = (w_1, \dots, w_n)$ are the corresponding sampling weights.

In the statistical software R, the user cannot continue working with the `quantile` function when the usage of sampling weights is required. Weighted quantiles are provided in different packages such as, among others, **Hmisc** (Harrell Jr et al., 2018), **laeken** (Alfons and Templ, 2013) and **survey** (Lumley, 2004). However, in contrast to the detailed descriptions of methodologies in the licensed software programs, it is not always clear from the package descriptions how the weights are incorporated. While Alfons and Templ (2013) describe the mathematical expression implemented in the **laeken** package that corresponds to WQ2, the documentations of the other packages only have a verbal description of how the weights are considered. The function `wtd.quantile` from the package **Hmisc** with type `quantile` is supposed to be a weighted version of the default definition used in the `quantile` function (Q3).

Since the original version of the Harrell-Davis estimator does not account for sampling weights, the weighted version is developed in this work in order to compare it to weighted versions of category 1 quantiles. The extension can be expressed by

$$WQ3_p = \sum_{i=1}^n W_{w_i, i} X_{(i)},$$

with weighting factor

$$\begin{aligned} W_{w_i, i} &= \frac{1}{\beta\{(\sum_i^n w_i + 1)p, (\sum_i^n w_i + 1)(1-p)\}} \int_{(i-1)/\sum_i^n w_i}^{s_i/\sum_i^n w_i} y^{(\sum_i^n w_i + 1)p-1} (1-y)^{(\sum_i^n w_i + 1)(1-p)-1} dy, \\ &= B_{s_i/\sum_i^n w_i}\{p(\sum_i^n w_i + 1), (1-p)(\sum_i^n w_i + 1)\} - B_{(i-1)/\sum_i^n w_i}\{p(\sum_i^n w_i + 1), (1-p)(\sum_i^n w_i + 1)\}, \end{aligned}$$

where $y = F(x)$ and $B_x(a, b)$ denotes the incomplete beta function and $s_i = s_{i-1} + w_i$ is the cumulated sum of ordered sampling weights corresponding to the order statistics. The R code can be requested from the author and a link to the code is provided in the online supplementary material of the paper.

3.2.3 Variance estimation

A point estimate that is based on survey data should always be reported with a measure of precision, e.g., the variance. However, many institutions that are responsible for official statistics do not provide the variance estimate along with the point estimate (see e.g., Deutsche Bundesbank, 2016; eurostat, 2018a). One reason is that the target group for the estimates does not use the information of the precision estimate. Another reason is that the variance estimates are

not as easy to obtain as the point estimates. Even though this work focuses on the evaluation of the point estimates of different quantile definitions, a small overview of variance estimation methods for quantiles and the availability of variance estimation methods in the software is provided.

As already mentioned, the distribution of the quantile $\hat{Q}_p = X_{(i+1)}$ with $i = \lfloor np \rfloor$ is shown to be asymptotically normal (see e.g., Cramér, 1946; Walker, 1968) and can be expressed by

$$\hat{Q}_p \sim N \left(Q_p, \frac{1}{f(Q_p)} \sqrt{\frac{p(1-p)}{n}} \right),$$

where Q_p is the corresponding population quantile and $f(\cdot)$ is the density of a sample with n values and one-dimensional distribution $F(\cdot)$.

However, this analytic expression for the variance is rarely used in practice since the underlying distribution of the sample $F(\cdot)$ has to be known or estimated nonparametrically. Furthermore, samples drawn using complex surveys cannot be considered in this expression. Instead, resampling methods like the jackknife method, balanced repeated replication (BRR), and the bootstrap are typically used for the variance estimation of non-linear statistics (Chatterjee, 2011). For sample quantiles, the delete-1 jackknife is known to be inconsistent and only modifications as proposed by Shao and Wu (1989) can be considered for the variance estimation. For the BRR method, consistency is shown by Shao and Wu (1992) and for the bootstrap by Babu (1986). However, several modifications of the latter are suggested (Shao, 1988; Cheung and Lee, 2005). Note that all of these studies define the quantile by the inverse of the ecdf such that the results do not need to be universal for all quantile definitions introduced in Section 3.2. For instance, Harrell and Davis (1982) propose a delete-1 jackknife for the Harrell-Davis estimator but they do not discuss its properties in detail. In the presence of complex surveys, Rust and Rao (1996) suggest the provision of replicate weights that allow the derivation of the variance estimates. This lowers the burden of the analyst to report these.

A general overview of computer software for variance estimation is given by Wolter (2007). Since this work focuses on the functions `EXAMINE`, `quantile`, `UNIVARIATE`, and `pctile`, we introduce the options for estimating the variance in these functions. The easiest way to receive standard errors and confidence intervals obtained by bootstrapping is offered by `SPSS` under the bootstrap option in function `EXAMINE`. The `quantile` function does not allow for any variance estimation. Thus, in `R` the user needs to use a function from another package, e.g., package `survey` (Lumley, 2004) or `emdi` (Kreutzmann et al., 2019). Another way would be to use a package for bootstrapping or a self-programmed bootstrap for the variance estimation. The `UNIVARIATE` procedure of `SAS`TM software allows for confidence intervals which, however, can only be estimated if no sampling weights are included. For the estimation of the variance of quantiles, the procedure `SURVEYMEANS` may be more appropriate. The user of the command `pctile` does not have an option for the variance estimation. One option in `Stata` is the `epctile` package (Kolenikov, 2017).

3.3 Comparison of quantile definitions

The studies that compare quantile definitions with regard to bias and accuracy measured by the mean squared error (MSE) consider different distributions. Parrish (1990) investigates the quantiles under a normal distribution. Dielmann et al. (1994) enlarges this work by analyzing skewed distributions that are common in business applications, such as the log-normal and the Pareto distribution. Also, the work of Schoonjans et al. (2011) is based on the normal and log-normal distribution. Harrell and Davis (1982) evaluate their quantile definition through different forms of the generalized lambda distribution. This work investigates the bias and accuracy of different quantile definitions for income and wealth data, since sample quantiles are often used to describe the distribution of these variables (see e.g., Deutsche Bundesbank, 2016; eurostat, 2018a). The first simulation study compares the quantile definitions $Q1 - Q6$ with the corresponding theoretical quantile. The theoretical distribution from which samples are drawn is the generalized beta distribution of the second kind (GB2) that is often used to describe income distributions (McDonald and Bordley, 1996). It is explained in Section 3.3.1 in more detail. The second simulation study is design-based for the comparison of the weighted quantile definitions $WQ1 - WQ3$ when the underlying variable is wealth. The data is obtained from the PHF (Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank, 2014) (see also Section 3.3.2).

In both studies, the relative bias (RB) and the relative root MSE (RRMSE) are measured for each quantile definition with different sample sizes based on $R = 10,000$ Monte-Carlo (MC) repetitions. The relative bias for each definition Q and each quantile level p is defined by

$$RB = \frac{1}{R} \sum_{r=1}^R \frac{(\hat{Q}_{pr} - Q_p)}{Q_p},$$

where \hat{Q}_{pr} is the quantile estimate in run r and Q_p is the value of the theoretical or population quantile.

The relative root mean squared error (RRMSE) can be expressed by

$$RRMSE = \sqrt{\frac{1}{R} \sum_{r=1}^R \left(\frac{(\hat{Q}_{pr} - Q_p)}{Q_p} \right)^2}.$$

While simulation results are presented in figures in Sections 3.3.1 and 3.3.2, the results are shown in tables in the online supplementary material of the paper.

From the theory and the mentioned literature the following hypotheses can be derived:

- The bias oscillates between bounds and the estimation of central quantiles is easier than the estimation of extreme ones (Okolewski and Rychlik, 2001). This oscillation property is independent of the sample size. It also follows that there is no monotone increase or decrease of the bias along quantile levels.
- With increasing sample size, the bias becomes symmetric and thus decreases in absolute terms (Okolewski and Rychlik, 2001). Also, the accuracy should increase with larger sample size for all quantile definitions (see Section 3.2.1 and 3.2.2 for the definitions).

Consequently, the quantile definitions show asymptotic equivalence (see e.g., Yoshizawa et al., 1985; Parrish, 1990; Vélez and Correa, 2014).

These results should also be seen in the following simulations. However, most of these hypotheses are derived from the theoretical properties of a simple quantile definition that is only based on one order statistic. In contrast, different quantile definitions are evaluated empirically in the following.

3.3.1 Simulation using income type data

In this simulation study, a theoretical distribution is used. Since the aim is to mimic an income distribution, the GB2 distribution is chosen with parameters following McDonald and Bordley (1996) and Graf and Nedyalkova (2014). The density of the GB2 distribution is defined as

$$f(x, \theta) = \frac{a}{b \mathbf{GB}(c, d)} \frac{(x/b)^{ac-1}}{(1 + (x/b)^a)^{c+d}}, \quad x \geq 0,$$

where $\mathbf{GB}(c, d)$ is the beta function and $\theta = (a, b, c, d)$ are the parameters.

Graf and Nedyalkova (2014) fit the GB2 distribution to the equalised disposable income obtained from the EU Statistics on Income and Living Conditions (EU-SILC) survey (eurostat, 2013b) using different estimation approaches. Based on the maximum pseudo-likelihood estimation using the full pseudo-loglikelihoods, they propose the parameters $a = 7.481, b = 16351, c = 0.400$ and $d = 0.468$ to closely approximate the German income distribution. For more information about modeling income via parametric distributions, we refer to McDonald (1984) and Kleiber and Kotz (2003). From this GB2 distribution, samples are drawn in every replication via simple random sampling with the different sample sizes $n = 5, n = 6, n = 10, n = 11, n = 15, n = 16, \dots, n = 55, n = 56$. Thus, an even and an uneven number of almost the same size is always used. Along the lines of Parrish (1990) and Dielmann et al. (1994), we use small sample sizes in order to investigate the small sample properties of the quantile estimators. For each sample, the quantile definitions $Q1 - Q6$ are evaluated and compared with the true theoretical quantile. Table 3.4 states the values of the theoretical quantiles for the considered quantile levels $q \in (0.1, 0.25, 0.5, 0.75, 0.9, 0.99)$ and Figure 3.1 shows the true density and the location of the theoretical quantiles.

Figure 3.2 shows the RB in % for the definitions $Q1 - Q6$ of the considered quantile levels and Figure 3.3 the corresponding RRMSE in %. For all quantile definitions and quantile levels, it can be concluded that the bias decreases and the accuracy increases with increasing sample size. The theoretical result that the bias is smaller for the central quantile levels is also confirmed. Furthermore, the larger the sample size, the closer the estimates of the different quantile definitions are to each other. However, the convergence of the quantile definitions

Table 3.4: Values of the 0.1th, 0.25th, 0.5th, 0.75th, 0.9th, and 0.99th quantile of the GB2 distribution.

0.1	0.25	0.5	0.75	0.9	0.99
8742.18	11956.40	15686.06	20264.09	26661.86	51568.42

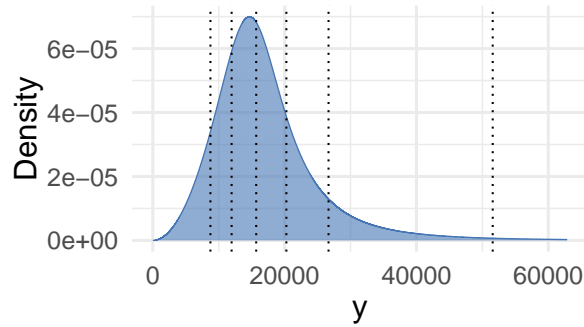


Figure 3.1: Density of the GB2 distribution (shown up to the 0.995th quantile). The dashed lines show the location of the 0.1th, 0.25th, 0.5th, 0.75th, 0.9th, and 0.99th quantile of the GB2 distribution (from left to right).

is slower for the extreme quantiles than for the central quantiles. For the median, the RB and the RRMSE is even the same for all quantile definitions of category 1, i.e., $Q1 - Q5$. The definition by Hazen (1914), $Q4$, is not always the least biased definition in contrast to the findings of Parrish (1990) and Schoonjans et al. (2011) when the underlying distribution is normal. However, it shows reasonably good results for all quantile levels. For the lower quantiles, 0.1 and 0.25, it can be seen that definition $Q5$, which is suggested by Hyndman and Fan (1996) as a common definition, performs very well ($|RB| < 2\%$) even for reasonably small sample sizes (from $n \geq 10$). Similarly, the definition $Q3$ (default in R) is especially good for the 0.75th quantile and definition $Q1$ (default in SASTM software and Stata) fluctuates around the 0 for the 0.9th quantile. The Harrell-Davis estimator is not the least biased estimator for any quantile level in this evaluation. Instead, it outperforms all quantile definitions with regard to accuracy for the quantile levels 0.1, 0.25 and 0.5 when the sample size is above 10. Furthermore, the definition $Q3$ shows the lowest RRMSE for almost all sample sizes for the higher quantiles, 0.75 and 0.9.

Since many researchers are especially interested in the upper end of the income distribution, results for the 0.99th quantile are shown in Figure 3.4. It is important to notice that the sample size needs to be larger in order to receive comparable relative biases to the other quantile levels. The best quantile definition for the 0.99th quantile with regard to the smallest RB is the definition by Hazen (1914) and the default of SASTM software and Stata. Regarding the smallest RRMSE, the default definition of the software R is best.

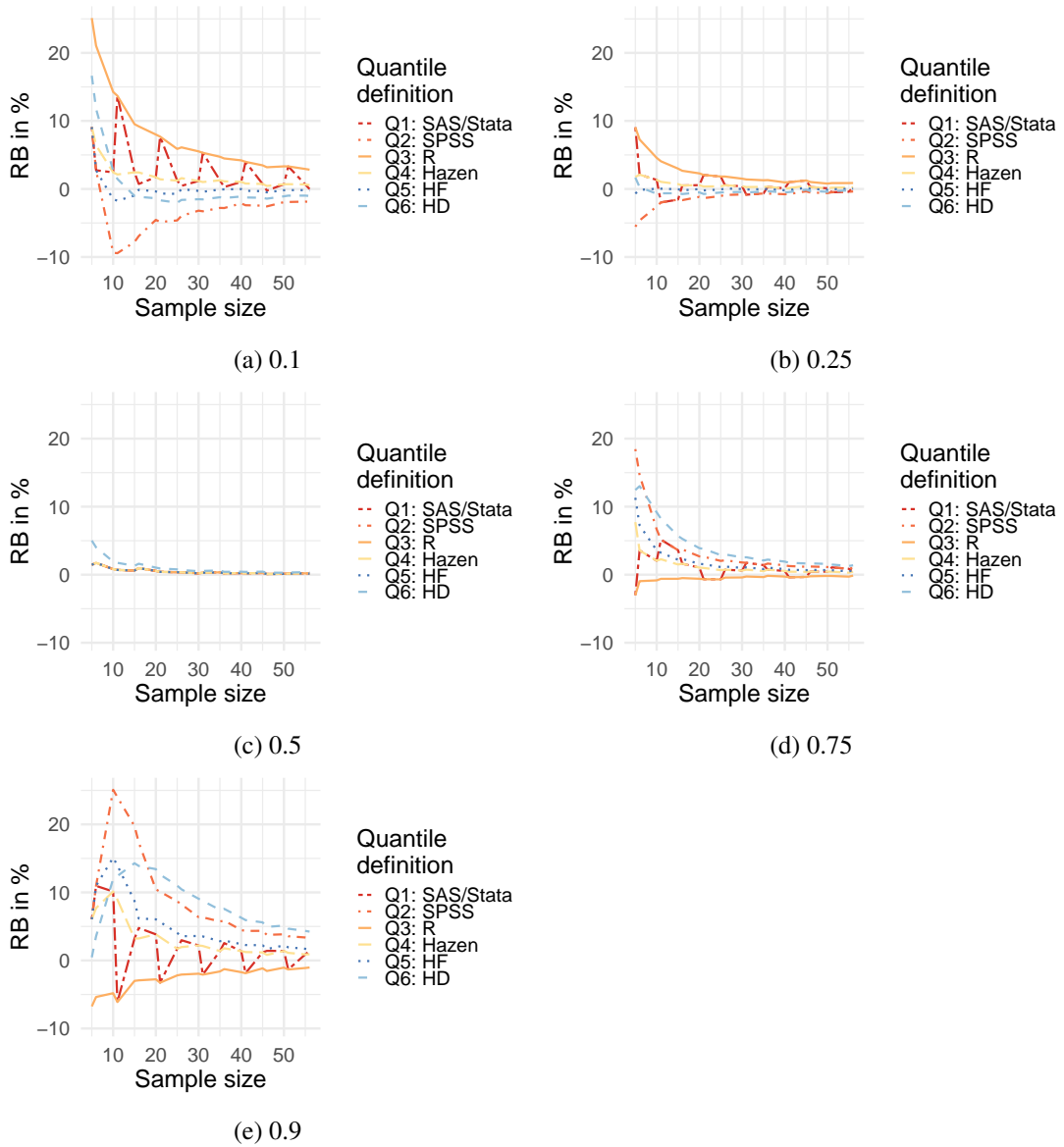


Figure 3.2: The plots show the RB in % for six different quantile definitions at the quantile levels 0.1 (a), 0.25 (b), 0.5 (c), 0.75 (d) and 0.9 (e) for different sample sizes from simulated data.

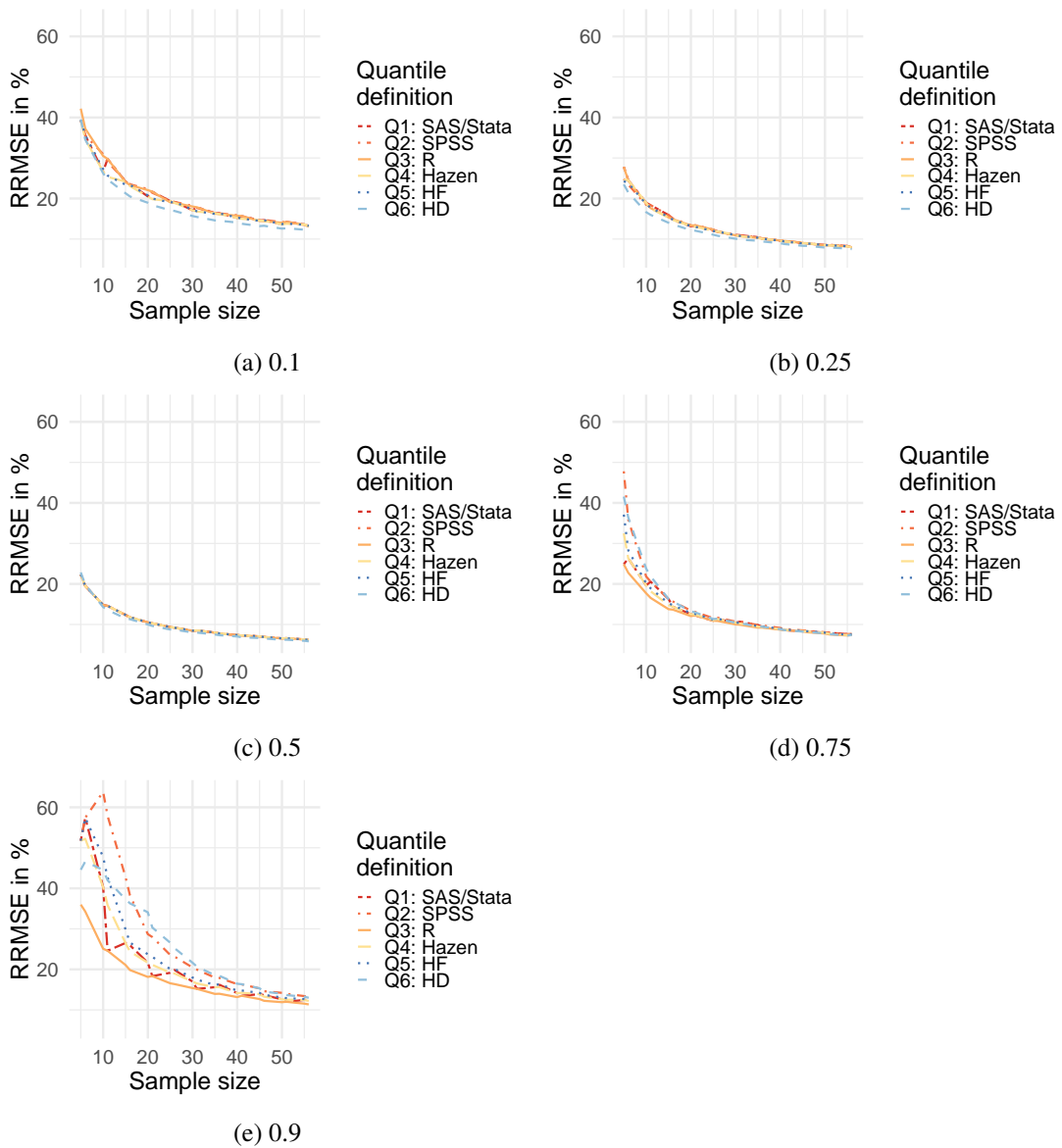


Figure 3.3: The plots show the RRMSE in % for six different quantile definitions at the quantile levels 0.1 (a), 0.25 (b), 0.5 (c), 0.75 (d) and 0.9 (e) for different sample sizes from simulated data.

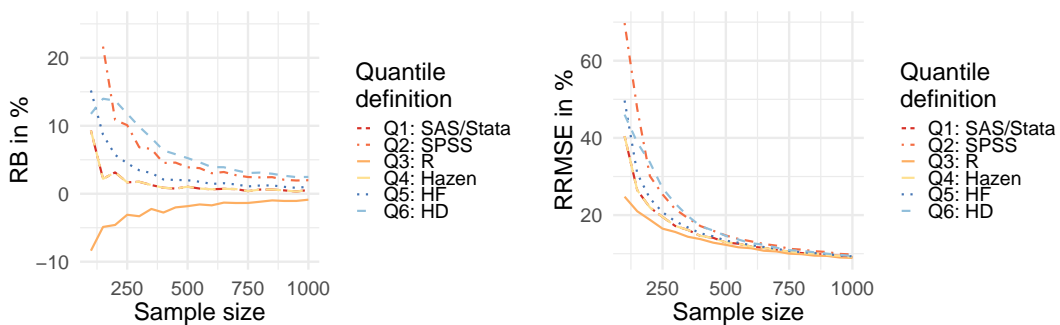
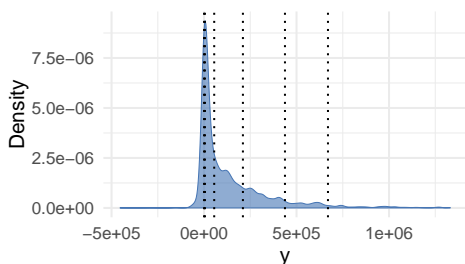


Figure 3.4: The plots show the RB (left) and the RRMSE (right) in % for six different quantile definitions at the quantile level 0.99 for different sample sizes from simulated data.



0.1	0.25	0.5	0.75	0.9	0.95
0	5000	55000	210200	437500	670000

Table 3.5: Values of the 0.1th, 0.25th, 0.5th, 0.75th, 0.9th, and 0.95th quantile of the synthetic net wealth distribution.

Figure 3.5: Density of the synthetic net wealth distribution based on Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014), Panel on Household Finances (PHF) 2014, own estimations (shown up to the 0.98th quantile). The dashed lines show the location of the 0.1th, 0.25th, 0.5th, 0.75th, 0.9th, and 0.95th quantile (from left to right).

3.3.2 Simulation using wealth data

For the design-based simulation study, the variable household net wealth from the PHF data is used (Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank, 2014). Net wealth is defined as the difference between assets and liabilities. Therefore, negative values for household net wealth are possible when liabilities exceed assets. In order to mimic the German wealth distribution, the sample is expanded by the final survey weights divided by 1,000 which leads to a synthetic population of around 37,000 households. The division is conducted to reduce the computational time of the simulation. Figure 3.5 shows the synthetic population and Table 3.5 the values of the quantile levels $q \in (0.1, 0.25, 0.5, 0.75, 0.9, 0.95)$. From this close-to-reality population, samples are drawn by stratified sampling in each simulation run. The sampling design is a simplification of the original sampling design of the PHF (Knerr et al., 2015). Particularly, the stratum variable that defines the four strata in the simulation – wealthy small municipalities, other small municipalities, wealthy street sections in large cities, and other street sections in large cities – is a combination of the two variables in the data that reflect the first two levels of the sampling design of the PHF. The sample sizes are determined by the percentages that are drawn from each strata. They equal 36, 44, 54, 90, 128, 166, 202, 239, 296, 556, 741, 926, 1113, 1299.

It should be noted that the empirical distribution of net wealth differs from the theoretical GB2 distribution in Section 3.3.1. It has a larger standard deviation which induces a larger bias at the same sample size (Okolewski and Rychlik, 2001). Therefore, larger sample sizes are chosen for this study. Furthermore, the empirical distribution is not strictly monotone. Observations are concentrated especially around 0 and low net wealth levels. Thus, the samples drawn from the synthetic population may be samples with ties and mid-quantiles, a quantile definition for discrete distributions, might be preferable (Genton et al., 2006; Ma et al., 2011). However, as already mentioned, variables like wealth are usually assumed to be continuous and common quantile estimators for continuous variables are used in practice. Therefore, the evaluation in this work follows these practices.

Figure 3.6 and Figure 3.7 show the results for the RB and the RRMSE for the quantiles $WQ1 - WQ3$ in %. These relative measures cannot be shown for the 0.1th quantile since the population quantile is equal to 0. For the other levels, it can be seen that all quantile definitions show comparable results. Thus, the inclusion of sampling weights has a similar effect as larger

sample sizes with regard to the equivalence of different quantile definitions. Furthermore, the evaluation confirms that the sample size needs to be larger than for the GB2 data in order to achieve a comparable level of RB and RRMSE. It is striking that the estimation of the 0.25th quantile works worse than the estimation of the quantile level 0.9 in relative terms for sample sizes below 741. Furthermore, the RB of the 0.9th quantile is comparable to the RB of the median for $n \leq 166$ and the estimation of the 0.75th is best in terms of the RB. Similar patterns can be seen for the RRMSE.

The findings can be explained with the sampling distribution of the quantile estimators and the oscillation of the bias. The sampling distribution of the quantile estimators evaluated on the wealth data is skewed especially for the smaller sample sizes in the simulation. This means that the 10,000 estimates obtained from the simulation are not distributed symmetrically around 0. Therefore, the relative measures favor the quantile levels with a large population value in terms of a low RB. With increasing sample sizes, the sampling distribution becomes more symmetric and the results are closer to what was expected. For instance, the RB of the 0.25th quantile is below the RB of the 0.9th quantile for $n \geq 741$. In order to complement the findings based on relative measures, we also present the bias in absolute terms, defined as $Bias = \frac{1}{R} \sum_{r=1}^R (\hat{Q}_{pr} - Q_p)$. Figure 3.8 shows the bias in € for different quantile levels holding the quantile definition fixed to *WQ2* and the sample size fixed to 551. It gives an exemplary illustration of the oscillation of the bias and shows that it can happen that even the absolute bias is smaller for, e.g., the 0.75th quantile than for the 0.5th quantile. However, the tendency that it is more difficult to estimate the extreme quantiles, and here especially the larger quantiles, can also be seen in this figure.

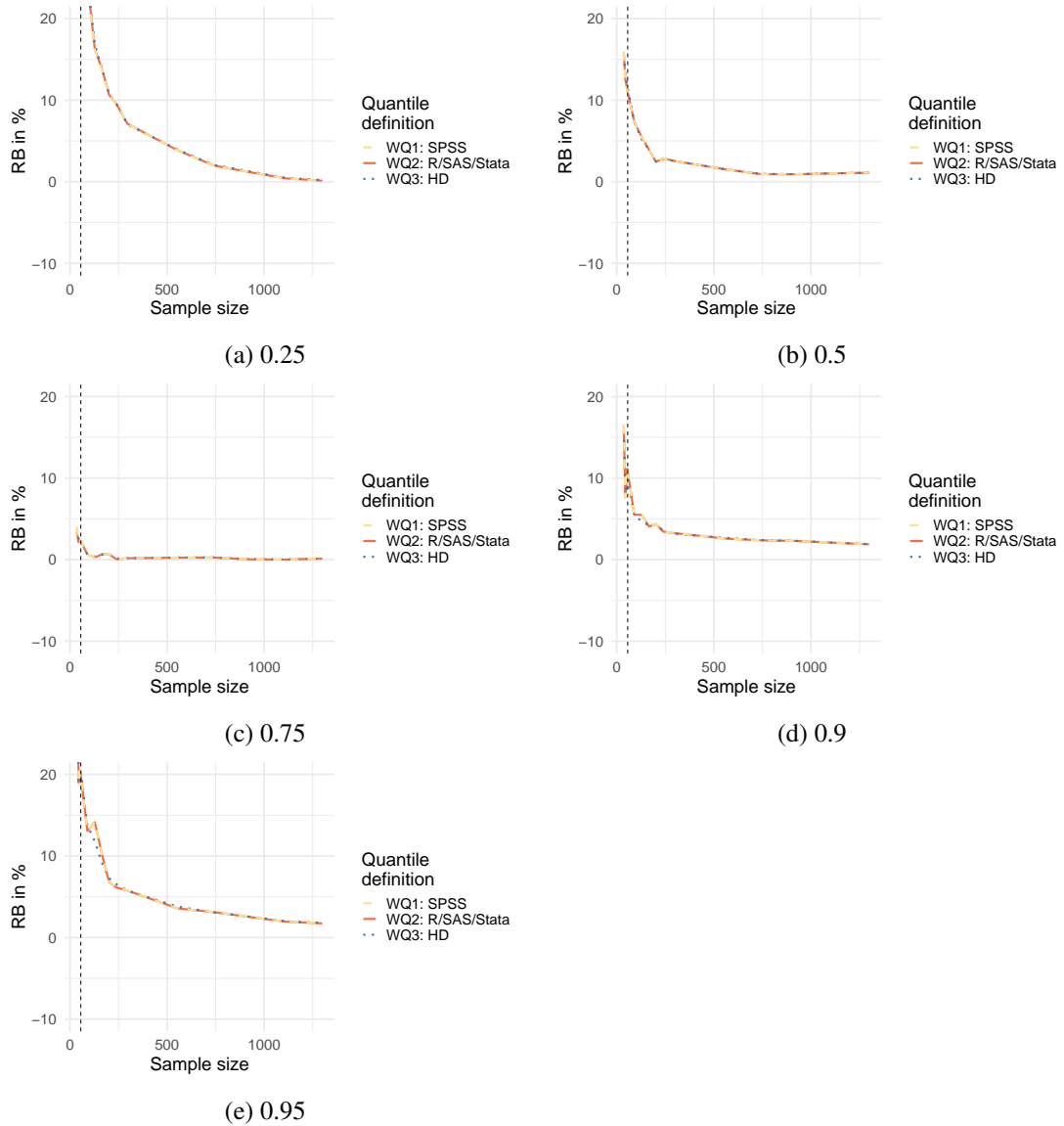


Figure 3.6: The plots show the RB in % for three different quantile definitions at the quantile levels 0.25 (a), 0.5 (b), 0.75 (c), 0.9 (d) and 0.95 (e) for different sample sizes. The dashed line indicates the sample size of 56 which is the largest sample size in the simulation based on income data. Furthermore, results that exceed an RB of 20% are not shown in the plot for consistency with the plots for income data.

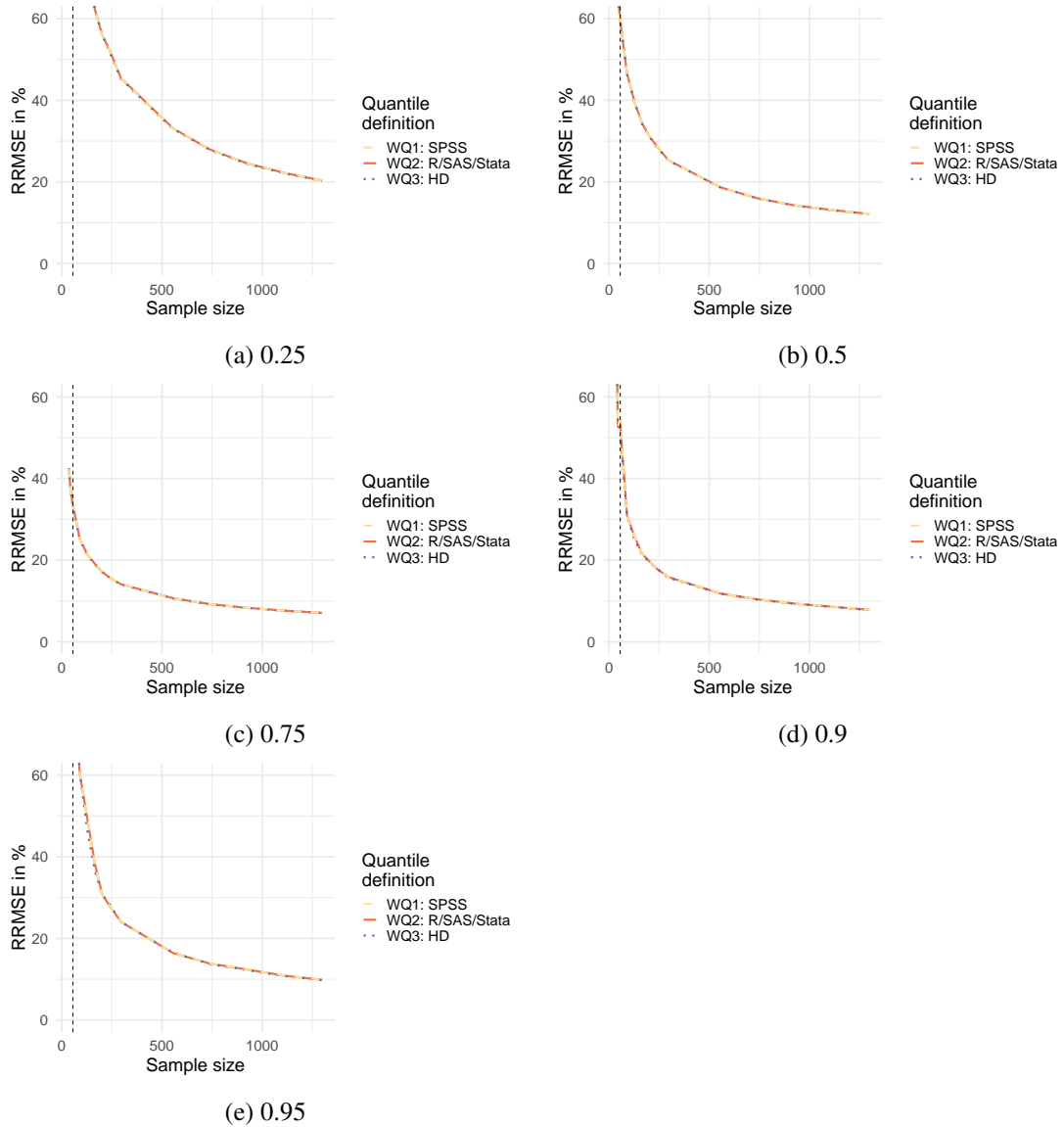


Figure 3.7: The plots show the RRMSE in % for three different quantile definitions at the quantile levels 0.25 (a), 0.5 (b), 0.75 (c), 0.9 (d) and 0.95 (d) for different sample sizes. The dashed line indicates the sample size of 56 which is the largest sample size in the simulation based on income data. Furthermore, results that exceed an RRMSE of 60% are not shown in the plot for consistency with the plots for income data.

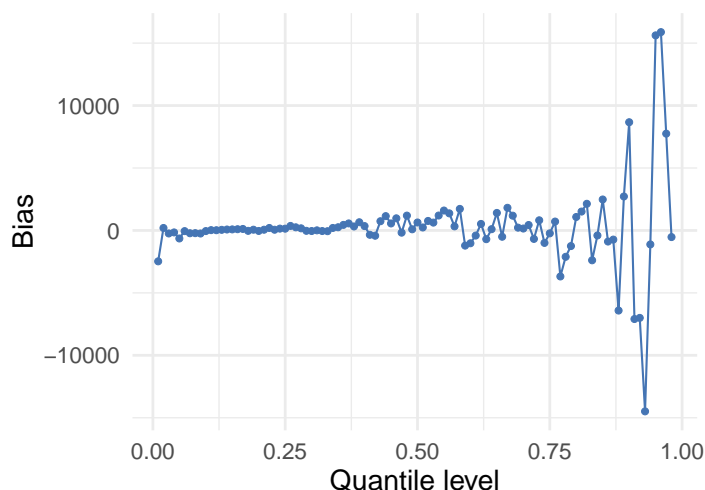


Figure 3.8: The plot shows the bias in € of the quantile definition $WQ2$ at different quantile levels when the sample size is fixed to 551.

3.4 Conclusion

This work compares quantile definitions implemented in the statistical software programs SPSS, R, SASTM software, and Stata by evaluating the performance on theoretical income and empirical wealth data. As a general result, it can be concluded that the performance of the quantile estimators differs depending on the quantile level, the sample size, and the type of distribution. This is in line with Parrish (1990) and Dielmann et al. (1994).

Comparing the software programs, the statistical software R provides the highest number of different quantile definitions. However, not all of these are also viable when sampling weights have to be considered. In this case, the choice of quantile definitions is limited in all software programs, except SPSS, that allows for sampling weights in all provided definitions. Furthermore, SPSS offers the simplest way for the estimation of the variance via a bootstrap. This work, however, only focuses on the standard options in the software programs. Due to the high number of different packages, especially in R, it is not claimed that all available possibilities are considered.

The comparison of the different quantile definitions leads to the following results. For the median, the different quantile definitions return almost equal estimates independent of the sample size. For the other quantile levels, differences between the definitions are obvious. However, with increasing sample size the results of the different quantile definitions converge for all quantile levels. Thus, the higher the sample size, the less important the choice of an appropriate quantile definition is. The comparison of the weighted quantiles reveals that all definitions are close to each other, regardless of the sample sizes chosen in the evaluation.

With regard to the different data types, the empirical evaluation of the quantile estimators shows almost unbiased estimates ($|RB| < 3\%$), even for small sample sizes below 30, when the underlying distribution is the theoretical GB2 distribution. For the empirical net wealth distribution, the sample size needs to be larger in order to reach the same level of RB. Thus, at which sample size and at which quantile level the quantile estimators perform well in terms of the bias and accuracy strongly depends on the type of distribution. Therefore, for the decision

in practical applications, it might be reasonable to evaluate different estimators via a design-based simulation, as proposed e.g., in Tzavidis et al. (2018), in order to get an indicator of the bias when the quantile is estimated for specific underlying distributions. Furthermore, the results should also be seen in the different contexts. While a bias of € 1000 might be negligible when wealth levels are of interest, it makes a huge difference for the analysis of income.

In addition to the bias, the analysis in this work reveals that the variance (as a component of the MSE) of the quantile estimators is larger, the smaller the sample size is. One way to handle this issue is with small area estimation (SAE) (Münnich et al., 2013). While many studies in SAE focus on the estimation of means or ratios (Marchetti et al., 2016; Schmid et al., 2017), methodology is also developed for other indicators including the median (Datta et al., 2002; Bell et al., 2016). Consequently, the application of SAE for quantile levels other than the median could be of interest in future research.

3.5 Supplementary material

The additional online material contains a) a PDF-file with code examples and outputs based on synthetic data in the different software programs, b) synthetic data and the code used to obtain the results in the aforementioned PDF-file, and c) tables with results from the simulation studies.

Acknowledgments

I gratefully acknowledge support by the German Research Foundation within the project QUESSAMI (281573942) and by the MIUR-DAAD Joint Mobility Program (57265468). This work uses data from the Deutsche Bundesbank Panel on Household Finances. The results published and the related observations and analysis may not correspond to results or analysis of the data producers.

Appendix B

B.1 Description of the Appendix

For the sake of completeness, the expressions of quantile estimators that are introduced in Table 3.2 but not mentioned in the text are shown in this Appendix. Furthermore, the six properties that are used by Hyndman and Fan (1996) are summarized.

B.1.1 Inverse of the empirical cumulative distribution function

Dielmann et al. (1994) states that this quantile estimator is neither mean nor median unbiased. For a further discussion of its properties, we refer to Juritz et al. (1983)

$$Q_p = \begin{cases} X_1 & \text{if } p = 0; \\ X_{(i)} & \text{if } 0 < p \leq 1 \text{ and } g = 0; \\ X_{(i+1)} & \text{if } 0 < p \leq 1 \text{ and } g \neq 0, \end{cases}$$

where $i = \lfloor np \rfloor$ and $g = np - i$.

B.1.2 Observation closest to np

This definition crucially depends on the rounding. While in R and SAS the rounding takes place to the next even integer, the definition in SPSS differs from the one below since it uses simple rounding.

$$Q_p = \begin{cases} X_{(1)} & \text{if } p \leq \frac{0.5}{n}; \\ X_{(i)} & \text{if } \frac{0.5}{n} < p \leq 1, \quad i \text{ is even and } g = 0; \\ X_{(i+1)} & \text{if } \frac{0.5}{n} < p \leq 1, \quad i \text{ is odd and } g \neq 0, \end{cases}$$

where $i = \lfloor np \rfloor$ and $g = np - 0.5 - i$.

B.1.3 Linear interpolation of the empirical distribution function

This definition is proposed by Parzen (1979).

$$Q_p = \begin{cases} X_{(1)} & \text{if } p < \frac{1}{n}; \\ (1 - \gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{1}{n} \leq p < 1; \\ X_{(n)} & \text{if } p=1, \end{cases}$$

where $i = \lfloor np_k \rfloor$, $p_k = \frac{np}{n}$, $\gamma = np_k - i$.

B.1.4 Approximation to $F(E(X_k))$ for the normal distribution

This definition is especially preferable when the underlying distribution is normal (Blom, 1958). Thus, it is often used for normal quantile-quantile plots.

$$Q_p = \begin{cases} X_{(1)} & \text{if } p < \frac{5/8}{n+1/4}; \\ (1 - \gamma)X_{(i)} + \gamma X_{(i+1)} & \text{if } \frac{5/8}{n+1/4} \leq p < \frac{n-3/8}{n+1/4}; \\ X_{(n)} & \text{if } p \geq \frac{n-3/8}{n+1/4}, \end{cases}$$

where $i = \lfloor np_k + \frac{p}{4} + \frac{3}{8} \rfloor$, $p_k = \frac{(np + \frac{p}{4} + \frac{3}{8}) - \frac{3}{8}}{n + \frac{1}{4}}$, $\gamma = np_k + \frac{p}{4} + \frac{3}{8} - i$.

B.1.5 Six desirable properties for sample quantile

Table B.1: Replication of Table 1 in Hyndman and Fan (1996) that shows their definition of six desirable properties for a sample quantile. For more information about the properties it is referred to Hyndman and Fan (1996).

P1: $Q(p)$ is continuous.
P2: $\text{Freq}(X_{(i)} \leq Q(p)) \geq pn$
P3: $\text{Freq}(X_{(i)} \leq Q(p)) = \text{Freq}(X_{(i)} \geq Q(1 - p))$
P4: Where $Q^{-1}(x)$ is uniquely defined, $Q^{-1}(X_{(i)}) + Q^{-1}(X_{(n-i+1)})$ for $i = 1, \dots, n$
P5: Where $Q^{-1}(x)$ is uniquely defined, $Q^{-1}(X_{(1)}) > 0$ and $Q^{-1}(X_{(n)}) < 1$
P6: $Q(0.5)$ is equal to the sample median defined by $[X_{(l)} + X_{(l+1)}] / 2$ if $n = 2l$ $X_{(l+1)}$ if $n = 2l + 1$

Supplementary material B

SPSS: Default (unweighted) quantile definition - Haverage (Q2 in the paper)

```
SPSS> EXAMINE VARIABLES=income
+       /PLOT NONE
+       /PERCENTILES(5,10,25,50,75,90,95) HAVERAGE
+       /STATISTICS NONE
+       /MISSING LISTWISE
+       /NOTOTAL.
```

		Percentile						
		Percentile						
		5	10	25	50	75	90	95
Gewichtetes Mittel (Definition 1)	income	5625,4465	8426,2040	12783,8325	15863,6750	17089,3875	27980,7530	40258,9375
Tukey-Angelpunkte	income			13106,0100	15863,6750	16849,3100		

SPSS: Default weighted quantile definition - Haverage (WQ1 in the paper)

```
SPSS> WEIGHT BY weights.
SPSS> EXAMINE VARIABLES=income
+       /PLOT NONE
+       /PERCENTILES(5,10,25,50,75,90,95) HAVERAGE
+       /STATISTICS NONE
+       /MISSING LISTWISE
+       /NOTOTAL.
```

		Percentile						
		Percentile						
		5	10	25	50	75	90	95
Gewichtetes Mittel (Definition 1)	income	5801,3200	9689,6100	11486,0200	14137,2600	16556,7500	17809,6200	19212,8500
Tukey-Angelpunkte	income			11486,0200	14137,2600	16556,7500		

R: Default (unweighted) quantile definition - type = 7 (Q3 in the paper)

```
R> quantile(data$income, probs = c(0, 0.1, 0.25, 0.5, 0.75,
+ 0.9, 1))

      0%      10%      25%      50%      75%      90%
5410.490 8433.516 13249.170 15863.675 16841.675 25146.777
      100%
53970.630
```

R: Hazen (unweighted) quantile definition - type = 5 (Q4 in the paper)

```
R> quantile(data$income, type = 5, probs = c(0, 0.1, 0.25,
+ 0.5, 0.75, 0.9, 1))

      0%      10%      25%      50%      75%      90%      100%
5410.49  8429.86 13106.01 15863.67 16849.31 26563.76 53970.63
```

R: Hyndman and Fan (unweighted) quantile definition - type = 8 (Q5 in the paper)

```
R> quantile(data$income, type = 8, probs = c(0, 0.1, 0.25,
+ 0.5, 0.75, 0.9, 1))

      0%      10%      25%      50%      75%      90%
5410.490 8428.641 12998.618 15863.675 16929.336 27036.094
      100%
53970.630
```

R: Harrell-Davis (unweighted) quantile definition, package Hmisc (Q6 in the paper)

```
R> install.packages("Hmisc")
R> library(Hmisc)
R> hdquantile(data$income, probs = c(0, 0.1, 0.25, 0.5, 0.75,
+ 0.9, 1))

      0.00      0.10      0.25      0.50      0.75      0.90
5410.490 8163.496 12459.966 15683.526 17788.450 27209.626
      1.00
53970.630
```

R: Weighted quantile in R, e.g., package laeken (WQ2 in the paper)

```
R> install.packages("laeken")
R> library(laeken)
R> weightedQuantile(data$income, data$weights,
+ probs = c(0, 0.1, 0.25, 0.5, 0.75, 0.9, 1))

[1] 5410.49 9689.61 11486.02 14137.26 16556.75 17809.62
[2] 53970.63
```

R: Weighted Harrell-Davis quantile definition available from github <https://github.com/akreutzmann/whdquantile/blob/master/whdquantile.R> (WQ3 in the paper)

```
R> whdquantile(data$income, smp_weight = data$weights,
+   probs = c(0, 0.1, 0.25, 0.5, 0.75, 0.9, 1))

      0.00      0.10      0.25      0.50      0.75      0.90      1.00
5410.49  9689.61 11486.02 14125.10 16556.75 17809.62 53970.63
```

SAS: Default (unweighted) quantile definition - 5 (Q1 in the paper)

```
SAS> title 'Default (unweighted) quantile definition';
+   ods select Quantiles;
+   proc univariate data=Data;
+   var income;
+   output out=percentiles_unweighted pctlpre=P_
+   pctlpts= 10 to 90 by 10;
+   run;
```

Default (unweighted) quantile definition

Die Prozedur UNIVARIATE
Variable: income

Quantile (Definition 5)	
Level	Quantil
100% Max	53970.63
99%	53970.63
95%	29040.28
90%	26563.77
75% Q3	16849.31
50% Median	15863.68
25% Q1	13106.01
10%	8429.86
5%	5801.32
1%	5410.49
0% Min	5410.49

SAS: Default weighted quantile definition - 5 (WQ2 in the paper)

```
SAS> title 'Default weighted quantile definition';
+   ods select Quantiles;
+   proc univariate data=Data;
+   var income;
+   weight weights;
+   output out=percentiles_weighted pctlpre=P_
+   pctlpts= 10 to 90 by 10;
+   run;
```

Default weighted quantile definition

Die Prozedur UNIVARIATE
Variable: income
Gewichtung: weights

Gewichtete Quantile	
Level	Quantil
100% Max	53970.63
99%	28335.00
95%	19212.85
90%	17809.62
75% Q3	16556.75
50% Median	14137.26
25% Q1	11486.02
10%	9689.61
5%	5801.32
1%	5410.49
0% Min	5410.49

Stata: Default (unweighted) quantile definition - default (Q1 in the paper)

```
Stata> pctlile quant_unweighted = income, nq(10) genp(percent)
Stata> list percent quant_unweighted in 1/10
```

	percent	quant_~d
1.	10	8429.859
2.	20	11651.66
3.	30	13747.46
4.	40	14534.3
5.	50	15863.67
6.	60	16570.54
7.	70	16787.43
8.	80	18034.66
9.	90	26563.77
10.	.	.

Stata: Default weighted quantile definition - default (WQ2 in the paper)

```
Stata> pctlile quant_weighted = income [w = weights], nq(10)
Stata> list percent quant_weighted in 1/10
```

	percent	quant_w~
1.	10	9689.61
2.	20	11486.02
3.	30	11817.3
4.	40	13106.01
5.	50	14137.26
6.	60	16063.72
7.	70	16556.75
8.	80	16706.03
9.	90	17809.62
10.	.	.

Default definitions of the other programs in R

SPSS default - type 6 (Q2 in the paper)

```
R> quantile(data$income, type = 6, probs = c(0, 0.1, 0.25, 0.5,
+ 0.75, 0.9, 1))
```

```
      0%      10%      25%      50%      75%      90%
5410.490 8426.204 12783.833 15863.675 17089.388 27980.753
      100%
53970.630
```

SAS/Stata default - type 2 (Q1 in the paper)

```
R> quantile(data$income, type = 2, probs = c(0, 0.1, 0.25, 0.5,
+ 0.75, 0.9, 1))
```

```
      0%      10%      25%      50%      75%      90%      100%
5410.49  8429.86 13106.01 15863.67 16849.31 26563.76 53970.63
```


Table B.1: The table shows the RB in % for the quantile levels 0.1, 0.25 and, 0.5 for six different quantile definitions evaluated on the GB2 distribution described in Section 3.3.1.

n	0.1						0.25						0.5					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
5	9.09	9.09	25.12	9.09	9.09	16.64	9.07	-5.54	9.07	1.76	-0.67	1.67	1.45	1.45	1.45	1.45	1.45	5.02
6	2.74	2.74	21.05	6.41	2.74	11.78	2.10	-4.63	7.27	2.10	-0.14	0.27	1.74	1.74	1.74	1.74	1.74	4.02
10	2.44	-9.37	14.26	2.44	-1.50	2.69	1.33	-2.57	4.57	1.33	0.03	-0.70	0.82	0.82	0.82	0.82	0.82	1.92
11	13.69	-9.44	13.69	2.12	-1.73	1.47	-1.98	-1.98	4.06	1.04	0.04	-0.64	0.70	0.70	0.70	0.70	0.70	1.75
15	2.44	-7.80	9.53	2.44	-0.97	-1.02	-1.57	-1.57	2.99	0.71	-0.05	-0.63	0.62	0.62	0.62	0.62	0.62	1.31
16	0.73	-6.93	9.20	2.43	-0.12	-1.15	0.51	-1.64	2.67	0.51	-0.21	-0.75	0.97	0.97	0.97	0.97	0.97	1.61
20	1.73	-4.54	8.00	1.73	-0.36	-1.37	0.56	-1.16	2.29	0.56	-0.01	-0.50	0.57	0.57	0.57	0.57	0.57	1.04
21	7.69	-4.86	7.69	1.42	-0.68	-1.67	2.06	-1.35	2.06	0.36	-0.21	-0.74	0.42	0.42	0.42	0.42	0.42	0.92
25	1.23	-4.62	5.88	1.23	-0.72	-2.02	1.83	-1.04	1.83	0.39	-0.08	-0.52	0.37	0.37	0.37	0.37	0.37	0.78
26	0.46	-3.96	6.11	1.59	-0.03	-1.61	0.54	-0.88	1.82	0.54	0.06	-0.47	0.36	0.36	0.36	0.36	0.36	0.69
30	1.17	-3.19	5.53	1.17	-0.28	-1.51	0.40	-0.81	1.51	0.40	-0.01	-0.44	0.20	0.20	0.20	0.20	0.20	0.52
31	5.31	-3.28	5.31	1.02	-0.42	-1.53	-0.84	-0.84	1.40	0.28	-0.10	-0.50	0.29	0.29	0.29	0.29	0.29	0.56
35	1.27	-2.79	4.73	1.27	-0.08	-1.22	-0.66	-0.66	1.28	0.31	-0.02	-0.39	0.35	0.35	0.35	0.35	0.35	0.63
36	0.28	-2.80	4.48	1.12	-0.06	-1.33	0.31	-0.68	1.30	0.31	-0.02	-0.34	0.21	0.21	0.21	0.21	0.21	0.43
40	1.04	-2.13	4.20	1.04	-0.02	-1.11	0.10	-0.77	0.96	0.10	-0.19	-0.50	0.24	0.24	0.24	0.24	0.24	0.44
41	3.99	-2.39	3.99	0.80	-0.27	-1.25	1.00	-0.75	1.00	0.13	-0.16	-0.46	0.21	0.21	0.21	0.21	0.21	0.42
45	0.74	-2.45	3.47	0.74	-0.33	-1.30	1.25	-0.35	1.25	0.45	0.19	-0.10	0.26	0.26	0.26	0.26	0.26	0.45
46	-0.24	-2.56	3.19	0.45	-0.50	-1.43	0.28	-0.53	1.02	0.28	0.01	-0.30	0.10	0.10	0.10	0.10	0.10	0.30
50	0.69	-1.93	3.32	0.69	-0.18	-1.08	0.10	-0.61	0.79	0.10	-0.14	-0.41	0.13	0.13	0.13	0.13	0.13	0.29
51	3.32	-1.92	3.32	0.70	-0.17	-0.98	-0.48	-0.48	0.86	0.19	-0.04	-0.32	0.17	0.17	0.17	0.17	0.17	0.32
55	0.67	-1.86	2.94	0.67	-0.18	-0.98	-0.41	-0.41	0.86	0.23	0.01	-0.21	0.20	0.20	0.20	0.20	0.20	0.38
56	-0.01	-1.90	2.83	0.56	-0.22	-1.03	0.25	-0.37	0.88	0.25	0.04	-0.21	0.15	0.15	0.15	0.15	0.15	0.30

Table B.2: The table shows the RB in % for the quantile levels 0.75 and 0.9 for six different quantile definitions evaluated on the GB2 distribution described in Section 3.3.1.

n	0.75						0.9					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
5	-3.01	18.46	-3.01	7.73	11.30	12.45	6.05	6.05	-6.75	6.05	6.05	0.46
6	3.57	14.64	-0.95	3.57	7.26	13.02	10.96	10.96	-5.37	7.70	10.96	3.55
10	2.05	6.65	-0.83	2.05	3.59	9.16	10.14	25.08	-4.81	10.14	15.12	11.99
11	5.15	5.15	-0.60	2.27	3.23	8.23	-6.11	24.01	-6.11	8.95	13.97	12.42
15	3.60	3.60	-0.60	1.50	2.20	5.70	3.23	19.76	-3.00	3.23	8.74	14.29
16	1.61	3.72	-0.50	1.61	2.32	5.27	4.77	17.34	-2.92	3.23	6.17	13.89
20	1.05	2.69	-0.60	1.05	1.60	3.90	3.84	10.45	-2.77	3.84	6.04	13.43
21	-0.66	2.64	-0.66	0.99	1.54	3.79	-3.26	10.10	-3.26	3.42	5.65	12.62
25	-0.68	2.03	-0.68	0.67	1.12	2.88	1.73	8.69	-2.19	1.73	4.05	11.00
26	0.73	2.12	-0.42	0.73	1.20	2.98	2.98	8.24	-2.07	1.97	3.56	10.46
30	0.61	1.80	-0.40	0.61	1.00	2.54	2.21	6.37	-1.95	2.21	3.60	9.08
31	1.75	1.75	-0.26	0.74	1.08	2.56	-2.08	6.34	-2.08	2.13	3.53	8.75
35	1.42	1.42	-0.35	0.53	0.83	2.10	1.38	5.73	-1.62	1.38	2.83	7.44
36	0.74	1.64	-0.16	0.74	1.04	2.22	2.56	5.85	-1.27	1.79	2.92	7.59
40	0.55	1.37	-0.27	0.55	0.83	1.93	1.37	4.47	-1.72	1.37	2.40	6.28
41	-0.38	1.27	-0.38	0.45	0.72	1.75	-1.87	4.34	-1.87	1.24	2.27	5.94
45	-0.32	1.16	-0.32	0.42	0.67	1.65	1.19	4.35	-1.16	1.19	2.24	5.60
46	0.46	1.21	-0.23	0.46	0.71	1.67	1.41	3.76	-1.55	0.82	1.67	4.93
50	0.45	1.13	-0.18	0.45	0.67	1.59	1.40	3.85	-1.05	1.40	2.21	5.13
51	1.08	1.08	-0.18	0.45	0.66	1.50	-1.32	3.56	-1.32	1.12	1.94	4.67
55	0.87	0.87	-0.28	0.29	0.48	1.28	0.89	3.38	-1.11	0.89	1.72	4.35
56	0.45	1.03	-0.13	0.45	0.64	1.38	1.40	3.27	-1.03	0.92	1.61	4.24

Table B.3: The table shows the RRMSE in % for the quantile levels 0.1, 0.25 and, 0.5 for six different quantile definitions evaluated on the GB2 distribution described in Section 3.3.1.

n	0.1						0.25						0.5					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
5	39.34	39.34	42.18	39.34	39.34	39.52	27.81	24.82	27.81	24.61	24.27	23.48	22.32	22.32	22.32	22.32	22.32	22.85
6	35.57	35.57	37.24	34.64	35.57	34.55	24.52	23.42	24.39	24.52	23.81	21.48	19.50	19.50	19.50	19.50	19.50	19.81
10	26.65	30.97	30.50	26.65	27.22	26.23	18.94	18.45	18.71	18.94	18.62	16.63	14.79	14.79	14.79	14.79	14.79	14.29
11	29.77	29.02	29.77	25.32	25.66	24.70	18.37	18.37	17.63	17.53	17.71	15.96	14.59	14.59	14.59	14.59	14.59	13.58
15	24.19	24.38	24.09	24.19	23.42	21.30	15.92	15.92	15.35	15.32	15.45	14.02	12.36	12.36	12.36	12.36	12.36	11.49
16	23.57	23.51	23.31	23.07	23.35	20.51	14.69	15.08	14.89	14.69	14.76	13.60	11.77	11.77	11.77	11.77	11.77	11.30
20	20.52	22.33	22.08	20.52	20.76	18.96	13.16	13.39	13.36	13.16	13.19	12.19	10.52	10.52	10.52	10.52	10.52	9.99
21	21.68	21.41	21.68	19.83	19.98	18.36	13.38	13.09	13.38	13.03	13.00	12.06	10.29	10.29	10.29	10.29	10.29	9.64
25	19.37	19.45	19.06	19.37	19.04	17.10	12.32	12.03	12.32	12.01	11.98	11.08	9.36	9.36	9.36	9.36	9.36	8.76
26	19.23	19.16	18.94	18.93	19.13	16.92	11.89	11.77	11.74	11.89	11.81	10.80	9.19	9.19	9.19	9.19	9.19	8.72
30	17.10	18.18	18.02	17.10	17.25	15.68	11.02	10.93	10.92	11.02	10.96	10.08	8.46	8.46	8.46	8.46	8.46	8.02
31	17.84	17.68	17.84	16.77	16.85	15.44	10.93	10.93	10.70	10.71	10.76	9.86	8.53	8.53	8.53	8.53	8.53	8.00
35	16.42	16.37	16.27	16.42	16.19	14.60	10.46	10.46	10.29	10.29	10.33	9.58	7.99	7.99	7.99	7.99	7.99	7.52
36	16.39	16.34	16.13	16.17	16.33	14.54	10.05	10.17	10.12	10.05	10.07	9.40	7.77	7.77	7.77	7.77	7.77	7.35
40	15.22	15.88	15.77	15.22	15.31	14.06	9.57	9.68	9.62	9.57	9.59	8.96	7.37	7.37	7.37	7.37	7.37	7.04
41	15.68	15.58	15.68	15.00	15.05	13.77	9.53	9.40	9.53	9.38	9.37	8.76	7.28	7.28	7.28	7.28	7.28	6.83
45	14.62	14.65	14.46	14.62	14.48	13.19	9.06	8.93	9.06	8.93	8.91	8.34	7.05	7.05	7.05	7.05	7.05	6.65
46	14.66	14.72	14.41	14.49	14.64	13.27	8.94	8.90	8.88	8.94	8.91	8.31	6.89	6.89	6.89	6.89	6.89	6.54
50	13.61	14.15	14.02	13.61	13.69	12.57	8.52	8.49	8.46	8.52	8.50	7.90	6.57	6.57	6.57	6.57	6.57	6.22
51	14.23	14.21	14.23	13.76	13.81	12.67	8.55	8.55	8.44	8.44	8.46	7.90	6.60	6.60	6.60	6.60	6.60	6.23
55	13.62	13.60	13.46	13.62	13.51	12.31	8.28	8.28	8.19	8.19	8.21	7.67	6.30	6.30	6.30	6.30	6.30	5.97
56	13.28	13.28	13.08	13.15	13.26	11.99	7.93	7.99	7.98	7.93	7.94	7.51	6.14	6.14	6.14	6.14	6.14	5.84

Table B.4: The table shows the RRMSE in % for the quantile levels 0.75 and 0.9 for six different quantile definitions evaluated on the GB2 distribution described in Section 3.3.1.

n	0.75						0.9					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
5	24.85	47.80	24.85	32.49	37.11	41.55	51.75	51.75	35.98	51.75	51.75	44.58
6	26.31	35.83	22.75	26.31	28.32	36.40	57.42	57.42	34.29	52.23	57.42	46.61
10	19.63	21.84	17.70	19.63	20.12	23.79	40.32	63.81	25.09	40.32	47.75	44.55
11	20.57	20.57	16.56	18.19	18.91	21.87	24.59	57.87	24.59	35.87	42.67	41.98
15	16.30	16.30	13.78	14.78	15.24	16.30	26.52	42.83	21.09	26.52	29.91	37.32
16	14.38	15.67	13.65	14.38	14.75	15.66	25.95	38.28	19.84	24.41	26.60	36.29
20	12.54	13.42	12.07	12.54	12.79	13.24	21.64	28.69	18.14	21.64	23.72	34.08
21	12.23	13.08	12.23	12.44	12.61	12.96	18.34	27.98	18.34	21.12	23.09	30.33
25	10.92	11.56	10.92	11.07	11.20	11.35	19.16	23.57	16.57	19.16	20.09	26.58
26	11.34	11.67	10.87	11.34	11.41	11.41	19.86	23.14	16.36	18.97	20.08	25.21
30	10.42	10.70	10.02	10.42	10.49	10.43	16.99	20.47	15.39	16.99	17.98	21.64
31	10.69	10.69	9.92	10.22	10.35	10.32	15.18	19.94	15.18	16.52	17.47	20.66
35	9.87	9.87	9.22	9.47	9.58	9.45	15.61	17.97	13.95	15.61	16.13	18.56
36	9.40	9.75	9.22	9.40	9.50	9.45	16.21	17.88	13.99	15.63	16.33	18.42
40	8.85	9.13	8.71	8.85	8.93	8.82	14.13	16.36	13.16	14.13	14.75	16.45
41	8.60	8.89	8.60	8.67	8.73	8.59	13.50	16.29	13.50	14.23	14.79	16.34
45	8.31	8.54	8.31	8.36	8.40	8.24	13.88	15.29	12.67	13.88	14.19	15.32
46	8.31	8.45	8.08	8.31	8.34	8.14	13.65	14.61	12.24	13.27	13.71	14.50
50	7.95	8.07	7.77	7.95	7.98	7.80	12.61	14.21	11.91	12.61	13.06	14.00
51	7.96	7.96	7.53	7.70	7.78	7.61	12.01	13.96	12.01	12.52	12.91	13.66
55	7.61	7.61	7.28	7.41	7.47	7.28	12.46	13.41	11.55	12.46	12.67	13.25
56	7.50	7.68	7.40	7.50	7.55	7.39	12.51	13.20	11.35	12.20	12.56	13.00

Table B.5: The table shows the RB and RRMSE in % for the 0.99 quantile level and six different quantile definitions evaluated on the GB2 distribution described in Section 3.3.1.

n	RB in %						RRMSE in %					
	Q1	Q2	Q3	Q4	Q5	Q6	Q1	Q2	Q3	Q4	Q5	Q6
100	9.30	26.98	-8.37	9.30	15.19	11.76	40.46	69.76	24.78	40.46	49.68	46.03
150	2.20	21.66	-4.88	2.20	8.69	13.98	26.46	47.52	21.00	26.46	30.79	38.81
200	3.14	10.87	-4.59	3.14	5.72	13.71	21.98	30.11	18.79	21.98	24.32	33.66
250	1.68	10.09	-3.09	1.68	4.49	11.78	19.54	25.37	16.50	19.54	20.75	26.96
300	1.76	6.83	-3.31	1.76	3.45	9.86	17.19	21.54	15.63	17.19	18.40	23.00
350	1.28	6.60	-2.23	1.28	3.05	8.21	16.16	19.23	14.39	16.16	16.81	19.76
400	0.88	4.53	-2.76	0.88	2.10	6.37	14.59	17.15	13.77	14.59	15.28	17.15
450	0.76	4.58	-2.01	0.76	2.03	5.82	14.11	15.95	12.86	14.11	14.50	16.00
500	1.04	3.90	-1.83	1.04	1.99	5.25	12.91	14.72	12.30	12.91	13.40	14.62
550	0.77	3.77	-1.57	0.77	1.77	4.63	12.58	13.83	11.66	12.58	12.84	13.65
600	0.66	3.02	-1.70	0.66	1.45	3.92	11.85	13.20	11.42	11.85	12.21	12.73
650	0.73	3.20	-1.30	0.73	1.55	3.90	11.51	12.52	10.77	11.51	11.73	12.33
700	0.69	2.73	-1.36	0.69	1.37	3.44	10.96	12.08	10.59	10.96	11.26	11.55
750	0.39	2.46	-1.35	0.39	1.08	3.05	10.57	11.27	10.02	10.57	10.72	10.92
800	0.62	2.41	-1.17	0.62	1.21	3.11	10.16	11.06	9.88	10.16	10.40	10.75
850	0.62	2.44	-0.98	0.62	1.23	2.95	10.01	10.65	9.50	10.01	10.15	10.36
900	0.50	2.05	-1.06	0.50	1.02	2.65	9.69	10.43	9.43	9.69	9.89	10.01
950	0.33	1.95	-1.06	0.33	0.87	2.45	9.43	9.95	9.02	9.43	9.54	9.69
1000	0.56	1.99	-0.88	0.56	1.04	2.47	9.14	9.80	8.91	9.14	9.31	9.42

Table B.6: The table contains the RB in % for three different quantile definitions at the quantile levels 0.25, 0.5, 0.75, 0.9 and 0.95 evaluated on the synthetic population described in Section 3.3.2.

n	0.25			0.5			0.75			0.9			0.95		
	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3
36	62.73	61.24	62.64	15.88	15.52	15.88	3.43	3.77	4.05	16.47	15.75	16.47	105.22	105.22	105.22
44	44.65	46.85	44.65	12.55	13.29	12.55	2.16	2.62	2.16	7.64	7.97	7.64	19.00	19.00	19.00
54	37.23	37.09	37.23	11.24	11.73	11.24	2.41	2.39	2.41	11.14	10.32	11.14	20.34	20.47	20.34
90	24.52	24.82	24.52	7.32	7.35	7.32	0.62	0.65	0.62	5.52	5.47	5.52	13.07	13.73	13.07
128	16.28	16.87	16.17	5.49	5.21	5.47	0.27	0.36	0.27	5.49	4.70	5.49	14.15	11.79	14.15
166	13.76	13.74	13.71	3.89	3.93	3.89	0.66	0.60	0.66	4.06	4.18	4.07	10.12	9.10	10.12
202	10.64	10.72	10.63	2.44	2.46	2.44	0.64	0.66	0.65	4.37	4.34	4.37	6.81	7.26	6.81
239	9.55	9.55	9.55	2.87	2.80	2.87	0.05	0.13	0.06	3.44	3.54	3.46	6.14	6.54	6.19
296	7.11	7.02	7.10	2.55	2.54	2.55	0.17	0.19	0.18	3.25	3.18	3.26	5.76	5.74	5.78
556	3.84	3.92	3.83	1.53	1.54	1.53	0.25	0.26	0.25	2.59	2.68	2.60	3.57	3.79	3.59
741	1.99	2.05	1.99	0.93	0.96	0.93	0.26	0.27	0.26	2.37	2.40	2.37	3.12	3.09	3.13
926	1.18	1.26	1.17	0.92	0.93	0.92	0.04	0.05	0.05	2.29	2.30	2.30	2.52	2.56	2.54
1113	0.41	0.46	0.40	1.03	1.03	1.03	0.01	0.03	0.01	2.08	2.09	2.09	1.94	2.00	1.95
1299	0.14	0.21	0.13	1.10	1.10	1.10	0.13	0.14	0.14	1.88	1.91	1.89	1.74	1.81	1.75

Table B.7: The table contains the RRMSE in % for three different quantile definitions at the quantile levels 0.25, 0.5, 0.75, 0.9 and 0.95 evaluated on the synthetic population described in Section 3.3.2.

n	0.25			0.5			0.75			0.9			0.95		
	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3	WQ1	WQ2	WQ3
36	188.96	184.75	188.96	73.72	73.65	73.72	42.35	42.53	42.36	84.00	82.96	84.00	356.94	356.92	356.94
44	149.09	149.57	149.09	67.15	67.11	67.15	37.61	37.71	37.61	52.74	52.76	52.74	89.71	89.71	89.71
54	128.27	126.77	128.27	60.20	60.36	60.20	33.95	33.75	33.95	54.30	52.45	54.30	93.00	92.88	93.00
90	91.63	91.32	91.63	46.58	46.34	46.58	25.25	25.15	25.25	30.94	30.75	30.94	60.48	60.81	60.48
128	72.86	73.46	72.83	39.43	39.42	39.42	21.41	21.24	21.41	25.68	24.92	25.68	49.39	46.45	49.39
166	62.20	62.20	62.13	34.29	34.09	34.28	19.08	18.96	19.08	21.52	21.28	21.52	38.21	36.48	38.21
202	56.07	56.03	56.06	31.17	31.04	31.17	17.15	17.06	17.15	19.67	19.49	19.67	30.81	30.79	30.81
239	52.30	51.82	52.29	28.68	28.48	28.68	15.76	15.72	15.75	17.86	17.92	17.86	27.98	28.36	27.98
296	45.41	45.18	45.40	25.32	25.24	25.32	14.10	14.01	14.10	15.87	15.71	15.87	24.05	24.06	24.08
556	33.11	32.95	33.10	18.67	18.62	18.67	10.68	10.62	10.68	11.82	11.81	11.82	16.41	16.49	16.41
741	27.91	27.79	27.91	15.99	15.93	15.99	9.21	9.15	9.21	10.40	10.34	10.41	13.83	13.71	13.84
926	24.54	24.41	24.54	14.24	14.19	14.24	8.32	8.28	8.32	9.36	9.30	9.36	12.38	12.24	12.41
1113	22.18	22.06	22.18	13.13	13.06	13.13	7.62	7.58	7.62	8.59	8.54	8.60	10.85	10.74	10.87
1299	20.33	20.24	20.33	12.15	12.09	12.15	7.10	7.05	7.10	7.84	7.77	7.84	9.86	9.80	9.87

Chapter 4

The R package `emdi` for the estimation and mapping of regional disaggregated indicators

4.1 Introduction

In recent years an increased number of policy decisions has been based on statistical information derived from indicators estimated at disaggregated geographical levels using small area estimation methods. Clearly, the more detailed the information provided by official statistics estimates, the better the basis for targeted policies and evaluating intervention programs. The United Nations suggest further disaggregation of statistical indicators for monitoring the Sustainable Development Goals. National statistical institutes (NSI) and other organizations across the world have also recognized the potential of producing small area statistics and their use for informing policy decisions. Examples of NSI with well-developed programs in the production of small area statistics include the US Bureau of Census, the UK Office for National Statistics (ONS) and the Statistical Office of Italy (ISTAT). Although the term domain is more general as it may include non-geographic dimensions, the term small area estimation (SAE) is the established one. We shall follow the custom in this paper and use the terms area/geography and domain/aggregation interchangeably.

Without loss of generality in this paper we will assume that the primary data sources used to estimate statistical indicators are national socio-economic household sample surveys. Sample surveys are designed to provide estimates with acceptable precision at national and possibly sub-national levels but usually have insufficient sizes to allow for precise estimation at lower geographical levels. Therefore, direct estimation that relies only on the use of survey data can be unreliable or even not possible for domains that are not represented in the sample. In the absence of financial resources for boosting the sample size of surveys, using model-based methodologies can help to obtain reliable estimates for the target domains.

Model-based SAE methods (Pfeffermann, 2013; Rao and Molina, 2015; Tzavidis et al., 2018) work by using statistical models to link survey data, that are only available for a part of the target population, with administrative or census data that are available for the entire

population. Despite the wide range of SAE methods that have been proposed by academic researchers, these are so far applied only by a fairly small number of NSI or other practitioners. This gap between theoretical advances and applications may have several reasons one of which is the lack of suitable, user friendly statistical software. More precisely, software needs not only to be available but it also needs to simplify the application of the methods for the user. The R (R Core Team, 2018) package **emdi** (Kreutzmann et al., 2018) aims to improve the user experience by simplifying the estimation of small area indicators and corresponding precision estimates. Furthermore, the user benefits from support that extends beyond estimation in particular, evaluating, processing, and presenting the results.

Traditionally model-based SAE methods have been employed for estimating simple, linear indicators for example, means and totals using for example, mixed (random) effects models and empirical best linear unbiased predictors (EBLUP). Several software packages exist. In R, the package **JoSAE** (Breidenbach, 2015) includes functions for EBLUP using unit-level models. Functions in the package **hbsae** (Boonstra, 2012) enable the use of unit- and area-level models and can be estimated either by using restricted maximum likelihood (REML) or hierarchical Bayes methods. The package **BayesSAE** (Shi and with contributions from Peng Zhang, 2013) also allows for Bayesian methods. The **rsae** package by Schoch (2014) and package **sae-Robust** by Warnholz (2016a) provide functions for outlier robust small area estimation using unit- or area-level models. Gaussian area-level multinomial mixed-effects models for SAE can be done with the **mme** package (Lopez-Vizcaino et al., 2014). In addition, resources in R are available for Bayesian SAE from the BIAS (Bayesian methods for combining multiple individual and aggregate data sources) project (Gómez-Rubio et al., 2010) and from the package **SAE2** (Gómez-Rubio et al., 2008) that provides likelihood-based methods. In Stata, functions `xtmixed` and `gllamm` support the estimation of linear mixed models, which is a popular basis for model-based SAE. EBLUP can be derived using these functions (West et al., 2007). Similarly, `PROC MIXED` and `PROC IML` can be used for fitting unit- and area-level models in SAS as shown in Mukhopadhyay and McDowell (2011). Furthermore, several SAS macros for SAE are provided by the EURAREA (enhancing small area estimation techniques to meet European needs) project (EURAREA Consortium, 2004).

More recently widespread application of SAE methods involves the estimation of poverty and inequality indicators and distribution functions (The World Bank, 2007). In this case the use of methodologies for estimating means and totals is no longer appropriate since such indicators are complex, non-linear functions of the data. As an example, we refer to the Foster-Greer-Thorbecke indicators (Foster et al., 1984), the Gini coefficient (Gini, 1912) and the quantiles of the income distribution. Popular SAE approaches for estimating complex indicators include the empirical best predictor (EBP) (Molina and Rao, 2010), the World Bank method (Elbers et al., 2003) and the M-Quantile method (Chambers and Tzavidis, 2006; Tzavidis et al., 2010). Although in this paper we focus exclusively on software for implementing the EBP method (Molina and Rao, 2010), a future version of the package will include the M-Quantile and World Bank methods. The World Bank provides a free software for using the World Bank method called **PovMap** (The World Bank Group, 2013). However, this focuses exclusively on poverty mapping. Creating a more general open-source software can help to accelerate the up-

take of modern model-based methods. Currently, the best known package that also includes the EBP method is the R package **sae** (Molina and Marhuenda, 2015). Although the **sae** package implements a range of small area methods, it lacks the necessary functionality for supporting the user beyond estimation for example, for performing model diagnostic analyses, visualizing, and exporting the results for further processing. In contrast, **emdi** supports the user by providing more options and greater flexibility. In particular, package **emdi** offers the following attractive features that distinguishes it from the **sae** package and other R packages for SAE:

- The estimation functions return by default estimates for a set of predefined indicators, including the mean, the quantiles of the distribution of the response variable and poverty and inequality indicators. Additionally, self-defined indicators or indicators available from other packages can be included.
- The user can select the type of data transformation to be used in **emdi**. Data-driven transformation parameters are estimated automatically.
- In contrast to other packages that include only a parametric bootstrap for mean squared error (MSE) estimation, package **emdi** includes two bootstrap methods, a parametric bootstrap and a semi-parametric wild bootstrap (see Appendix C.1) for MSE estimation. Both incorporate the uncertainty due to the estimation of the transformation parameter. The use of wild bootstrap (Flachaire, 2005; Thai et al., 2013) protects the user against departures from the distributional assumptions of the nested error linear regression model. This offers additional protection against possible misspecification of the model assumptions.
- Customized parallel computing is offered for reducing the computational time associated with the use of bootstrap.
- Package **emdi** provides predefined functions for diagnostic analyses of the model assumptions. A mapping tool for plotting the estimated indicators enables the creation of high quality visualization. The output summarizing the most relevant results can be exported to ExcelTM and to OpenDocument Spreadsheets for presentation and reporting purposes.

The remainder of this paper is structured as follows. Section 4.2 gives information about the estimation methods that are included in the package. In Section 4.3 we present the data sets that we used for illustrating the use of the **emdi** package. Section 4.4 describes the core functionality of the package. Examples demonstrate the use of the methods for computing, assessing and presenting the estimates. Section 4.5 shows how users can extend the set of indicators to be estimated by including customized options and describes the parallelization features of the package. Finally, Section 4.6 discusses future potential extensions.

4.2 Statistical methodology

In order to obtain regionally disaggregated indicators, package **emdi** includes direct estimation and currently model-based estimation using the EBP approach by Molina and Rao (2010). The

Measurement	Indicator I_i	Expression	Range
Location	Mean _{i}	$\frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}$	\mathbb{R}
	Q _{i,q}	$F_i^{-1}(q) = \inf\{y_i \in \mathbb{R} : F_i(y_i) \geq q\}$	\mathbb{R}
Poverty	HCR _{i}	$\frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{I}(y_{ij} \leq z)$	$[0, 1]$
	PG _{i}	$\frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z} \right) \mathbf{I}(y_{ij} \leq z)$	$[0, 1]$
Inequality	Gini _{i}	$\frac{2 \sum_{j=1}^{N_i} j y_{ij}}{N_i \sum_{j=1}^{N_i} y_{ij}} - \frac{(N_i+1)}{N_i}$	$[0, 1]$
	QSR _{i}	$\frac{\sum_{j=1}^{N_i} \mathbf{I}(y_{ij} > Q_{i,0.8}) y_{ij}}{\sum_{j=1}^{N_i} \mathbf{I}(y_{ij} \leq Q_{i,0.2}) y_{ij}}$	\mathbb{R}

Table 4.1: List of predefined population indicators in **emdi**. Note that $F_i(y_i)$ denotes the empirical distribution function of the population in domain i and quantiles are generally defined for $q \in (0, 1)$. The predefined quantiles in **emdi** are $q \in (0.1, 0.25, 0.5, 0.75, 0.9)$.

predefined indicators returned by the estimation functions in **emdi** include the mean and quantiles Q_q (10%, 25%, 50%, 75%, 90%) of the target variable as well as non-linear indicators of the target variable. A widely used family of indicators measuring income deprivation and inequality is the Foster-Greer-Thorbecke (FGT) one (Foster et al., 1984). Package **emdi** includes the FGT measures of headcount ratio (HCR)₁ and poverty gap (PG). In order to compute the HCR and PG indicators one must use a threshold z , also known as poverty line. This line is a minimum level of income deemed adequate for living in a particular country and can be defined in terms of absolute or relative poverty. For instance, the international absolute poverty line is currently set to \$1.90 per day by the World Bank (The World Bank, 2017). Relative poverty means a low income relative to others in a particular country – for instance, below a percentage of the median income in that country. Another family of indicators of interest is the so-called Laeken indicators, endorsed by the European Council in Laeken, Belgium (Council of the European Union, 2001). Two examples of Laeken indicators that are well-known for measuring inequality are the Gini coefficient (Gini, 1912) and the income quintile share ratio (QSR) (eurostat, 2004). These two inequality indicators are also available in **emdi**. Therefore, in total **emdi** includes ten predefined indicators I_i – summarized in Table 4.1 – that are estimated at domain level i using a) direct estimation introduced in Section 4.2.1 and b) model-based estimation via the EBP method introduced in Section 4.2.2.

In the following sections the notation denotes by U a finite population of size N , partitioned into D domains U_1, U_2, \dots, U_D of sizes N_1, \dots, N_D , where $i = 1, \dots, D$ refers to an i th domain and $j = 1, \dots, N_i$ to the j th household/individual. From this population a random sample of size n is drawn. This leads to n_1, \dots, n_D observations in each domain. If n_i is equal to 0 the domain is not in the sample. The target variable is denoted by y_{ij} .

4.2.1 Direct estimation

Direct estimation relies on the use of sample data only. The definition of direct (point and variance) estimators in **emdi** follows Alfons and Tempel (2013). The mean and the quantiles help to describe the level and the distribution of a target variable. Especially for target variables

with a skewed distribution, quantiles can be more appropriate summary statistics than the mean, since these are robust to extreme values. Direct estimators of the mean and the quantiles are defined as follows,

$$\widehat{\text{Mean}}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

$$\widehat{Q}_{i,q} = \begin{cases} \frac{1}{2} (y_{ik} + y_{ik+1}) & \text{if } \sum_{j=1}^k w_{ij} = q \sum_{j=1}^{n_i} w_{ij}; \\ y_{ik+1} & \text{if } \sum_{j=1}^k w_{ij} \leq q \sum_{j=1}^{n_i} w_{ij} \leq \sum_{j=1}^{k+1} w_{ij}, \end{cases}$$

where w_{ij} denotes the sample weights and $q \in (0, 1)$ defines the corresponding quantile.

The FGT measures HCR and PG are estimated by package **emdi** as follows,

$$\widehat{\text{HCR}}_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} \mathbf{I}(y_{ij} \leq z),$$

$$\widehat{\text{PG}}_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} \left(\frac{z - y_{ij}}{z} \right) \mathbf{I}(y_{ij} \leq z),$$

where the indicator function $\mathbf{I}(\cdot)$ equals 1 if the condition is met and 0 otherwise. As already mentioned, for the computation of the HCR and PG indicators one must use a threshold z , also known as the poverty line. Package **laeken** (Alfons and Templ, 2013) uses relative poverty lines defined as 60% of median equivalized disposable income, which corresponds to the EU definition for poverty lines and thus in this case the HCR is the At-risk-of-poverty rate. In contrast, package **emdi** allows both for absolute and relative poverty lines and the user is free to set the poverty line. Therefore, the threshold can be given as an argument in **emdi** or, alternatively, the user can define an arbitrary function depending on the target variable and sampling weights. As a default, a relative threshold defined as 60% of the median of the target variable is used. The HCR describes the proportion of the population below the poverty line and the PG further takes into account how far, on average, this proportion falls below the threshold. Both indicators are between 0 and 1.

The two inequality indicators Gini and QSR are estimated in **emdi** by

$$\widehat{\text{Gini}}_i = \left[\frac{2 \sum_{j=1}^{n_i} \left(w_{ij} y_{ij} \sum_{k=1}^j w_{ik} \right) - \sum_{j=1}^{n_i} w_{ij}^2 y_{ij}}{\sum_{j=1}^{n_i} w_{ij} \sum_{j=1}^{n_i} w_{ij} y_{ij}} - 1 \right],$$

$$\widehat{\text{QSR}}_i = \frac{\sum_{j=1}^{n_i} \mathbf{I}(y_{ij} > Q_{i,0.8}) w_{ij} y_{ij}}{\sum_{j=1}^{n_i} \mathbf{I}(y_{ij} \leq Q_{i,0.2}) w_{ij} y_{ij}},$$

where $\mathbf{I}(\cdot)$ is an indicator function that equals 1 if the target variable is above the weighted 80% quantile or below the 20% quantile and 0 otherwise. The Gini coefficient is between 0 and 1, and the higher the value, the higher the inequality is. The extreme values of 0 and 1 indicate perfect equality and inequality, respectively. QSR is typically used when the target variable is income and in this case it is defined as the ratio of total income of the 20% richest households to the 20% poorest households. The higher the value of QSR, the higher the inequality is.

While variance estimation in package **laeken** (Alfons and Templ, 2013) is only available for

the poverty and inequality indicators, package **emdi** also provides a non-parametric bootstrap method (Alfons and Templ, 2013) for estimating the variance of estimates of the mean and the quantiles. The variance is, on the one hand, an important measure for measuring the precision of estimates. On the other hand, it is also important to compute the coefficient of variation (CV) which is one measure for showing the extent of the variability of the estimate. The CV is used, for instance, by NSI for quantifying the uncertainty associated with the estimates and is defined as follows,

$$CV = \frac{\sqrt{\widehat{MSE}(\hat{I}_i)}}{\hat{I}_i},$$

where \hat{I}_i is an estimate of an indicator I_i for domain i and $\widehat{MSE}(\hat{I}_i)$ is the corresponding mean squared error.

4.2.2 Model-based estimation

The implementation of the EBP method in package **emdi** is based on the theory described in Molina and Rao (2010) and Rao and Molina (2015). The underlying model is a unit-level mixed model also known in SAE literature as the nested error linear regression model (Battese et al., 1988). In its current implementation the EBP method is based on a two-level nested error linear regression model that includes a random area/domain-specific effect and a unit (household or individual)-level error term.

In addition to the notation above, here we assume that $\mathbf{X} = (\mathbf{x}_0, \dots, \mathbf{x}_p)^\top$ is the design matrix, containing p explanatory variables. The nested error linear regression model is defined by

$$T(y_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, D, \quad u_i \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad e_{ij} \stackrel{iid}{\sim} N(0, \sigma_e^2), \quad (4.1)$$

where T denotes a transformation of the target variable y_{ij} , \mathbf{x}_{ij} is a vector of unit-level auxiliary variables of dimension $(p+1) \times 1$, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients and u_i and e_{ij} denote the random area and unit-level error terms. The EBP approach works by using at least two data sources, namely a sample data set used to fit the nested error linear regression model and a population (e.g., census or administrative) data set used for predicting – under the model – synthetic values of the outcome for the entire population. Both data sources must share identically defined covariates but the target variable is only available in the sample data set.

Use of data transformations

Under model (4.1) we assume that the model error terms follow a Gaussian distribution. However, in certain applications – as is the case when analyzing economic variables – this assumption may be unrealistic. Package **emdi** includes the option of using a one-to-one transformation $T(y_{ij})$ of the target variable y_{ij} aiming to make the Gaussian assumptions more plausible. A logarithmic-type transformation is very often used in practice (Elbers et al., 2003; Molina and Rao, 2010). However, this is not necessarily the optimal transformation, for example, when the

model error terms do not follow exactly a log-normal distribution. In addition to a logarithmic transformation, package **emdi** allows the use of a data-driven Box-Cox transformation (Box and Cox, 1964). The Box-Cox transformation is denoted by

$$T(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y_{ij} + s) & \text{if } \lambda = 0, \end{cases} \quad (4.2)$$

where λ is an unknown transformation parameter and s denotes the shift parameter, which is a constant and chosen automatically such that $y_{ij} + s > 0$. A general algorithm for estimating the transformation parameter λ is the REML, which is described in detail in Rojas-Perilla et al. (2017). One advantage of using the Box-Cox transformation is that it includes the logarithmic and no transformation as cases for specific values of λ . Package **emdi** currently includes the following options: no transformation, logarithmic transformation and Box-Cox transformation.

The EBP method is implemented using the following algorithm:

1. For a given transformation obtain $T(y_{ij}) = y_{ij}^*$. If the user selects the Box-Cox transformation, the transformation parameter λ is automatically estimated by the **emdi** package.
2. Use the sample data to fit the nested error linear regression model and estimate θ denoted by $\hat{\theta} = (\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$. The variance components are estimated by REML using the function `lme` from the package **nlme** (Pinheiro et al., 2015). Also compute $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

3. For $l = 1, \dots, L$:

- (a) For in-sample domains (domains that are part of the sample data set), generate a synthetic population of the target variable by $y_{ij}^{*(l)} = \mathbf{x}_{ij}^\top \hat{\beta} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)}$, with $v_i^{(l)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i))$, $e_{ij}^{(l)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_e^2)$ and \hat{u}_i , the conditional expectation of u_i given \mathbf{y}_i^* .

For out-of-sample domains (domains with no data in the sample) the conditional expectation of u_i cannot be computed, hence for these domains generate a synthetic population by using $y_{ij}^{*(l)} = \mathbf{x}_{ij}^\top \hat{\beta} + v_i^{(l)} + e_{ij}^{(l)}$, with $v_i^{(l)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_u^2)$, $e_{ij}^{(l)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_e^2)$.

For additional details we refer to Molina and Rao (2010).

- (b) Back-transform to the original scale $\mathbf{y}_i^{(l)} = T^{-1}(\mathbf{y}_i^{*(l)})$ and calculate the target indicator $I_i^{(l)}(\mathbf{y}_i^{(l)})$ in each domain. Note that $I_i^{(l)}$ is used here as a generic notation for any indicator of interest.

4. Compute the final estimates by taking the mean over the L Monte Carlo simulations in each domain, $\hat{I}_i^{EBP} = 1/L \sum_{l=1}^L I_i^{(l)}(\mathbf{y}_i^{(l)})$.

The **emdi** package fits the nested error linear regression model by using the **nlme** package and currently does not permit the use of an alternative package for example **lme4** (Bates et al., 2015). The reason for this choice is that in future developments of **emdi** we plan to allow for more complex covariance structures for the unit-level error term and the random effect for example, allowing for spatially correlated errors (Pratesi and Salvati, 2009; Schmid et al., 2016).

To the best of our knowledge, the **nlme** package offers sufficient flexibility for incorporating such models.

Measuring the uncertainty of the EBP estimates is done by using bootstrap methods. Here the uncertainty is quantified by the MSE. Package **emdi** includes two bootstrap schemes. One is parametric bootstrap under model 4.1 following Molina and Rao (2010), which additionally includes the uncertainty due to the estimation of the transformation parameter (Rojas-Perilla et al., 2017). Using an appropriate transformation often mitigates the departures from normality. However, even after transformations, departures from normality may still exist in particular for the unit-level error term. For this reason, **emdi** also includes a variation of semi-parametric wild bootstrap (Flachaire, 2005; Thai et al., 2013; Rojas-Perilla et al., 2017) to protect against departures from the model assumptions. The semi-parametric wild bootstrap is presented in detail in Appendix C.1. A simulation study comparing the performance of both MSE estimators is presented in Rojas-Perilla et al. (2017). Since the bootstrap schemes presented here are computationally intensive, **emdi** includes an option for parallelization that is described in detail in Section 4.5.2.

4.3 Data sets

The main idea of SAE is to combine multiple data sources. Typically, one data set is obtained from survey data at unit-level and the other one from census or administrative/register data. The target variable is available in the survey but not in the census data. The administrative data contains explanatory variables that are potentially correlated with the target variable and hence they can be used to assist the estimation. Depending on the model type and the indicator of interest, census information is needed at the unit-level, i.e., information is available for every unit in each domain, or it is required at the area-level which means that aggregated data for each domain is given. If the user is interested in estimating non-linear functions of the target variable (like indicators discussed in Section 4.2), then access to unit-level census data is needed. As the EBP approach in package **emdi** is suitable for estimating non-linear indicators, one population data set (`eusilcA_pop`) and one survey data set (`eusilcA_smp`) are provided at the household level such that the method can be illustrated. The two data sets are based on the use of `eusilcP` from the package **simFrame** (Alfons et al., 2010). This data set is a simulated close-to-reality version of the European Union Statistics on Income and Living Conditions (EU-SILC) in Austria from 2006. Austria is a federal republic in Central Europe made up of nine states and 94 districts (79 districts headed by commissions and 15 statutory cities) with a total population of about 8.8 million in 2018. The original EU-SILC data is obtained from an annual household survey that is nowadays conducted in all EU member states and six other European countries and enables the analysis of income, socio-demographic factors and living conditions.

For practical reasons, we need to modify the `eusilcP` data set used in package **simFrame**. Due to confidentiality constraints the lowest geographical level in this data set includes the nine states and identifiers for lower regional levels, like the 94 districts, are not included. However, in the context of SAE the interest is on lower geographical levels like districts or

municipalities. Therefore, we assigned households to Austrian districts for illustrating the methodology better. The modified synthetic population is called `eusilcA_pop`. The assignment is based on two criteria available from external sources: a) the population sizes at state and district level and b) the income level in each district. From the last register-based census in 2011 the population sizes in each district and in each state are known and publicly available (Statistik Austria, 2013). We defined the district population sizes in relation to the state population sizes in the `eusilcA_pop` data set such that their population ratios mimic the *true* ratios in Austria. Furthermore, the Austrian Chamber of Commerce published a ranking of the districts within the states based on the net per capita income (Wirtschaftskammer Österreich, 2017). Based on this ranking we assigned households to districts such that the ordering of the districts within states is maintained. One drawback of the population data set is the small number of households in some districts. For instance, the number of households is only 5 in Rust (Stadt). This is, however, partly due to the fact that it is also in reality a really small district with only 1896 inhabitants (Statistik Austria, 2013). Although the `eusilcA_pop` data set in **emdi** mimics some real characteristics in Austria, it is a synthetic population data set for demonstrating the functionality of the package and conclusions about the levels of inequality and poverty in the Austrian districts observed from this data are not official estimates. The documented complete code for the assignment of the households to the districts is available as supplementary file at the Journal of Statistical Software along with our article.

The target variable in the example is the equivalized household income (`eqIncome`), which is defined as the total household disposable income divided by the equivalized household size determined by the modified scale of the Organisation for Economic Co-operation and Development (OECD) (Hagenaars et al., 1994). Thus, the indicators in our illustration describe the distribution of income, poverty and inequality similarly to the analysis in Alfons and Templ (2013). The remaining variables in `eusilcA_pop` are variables that identify the regional levels (`state` and `district`) and auxiliary variables that can be used for modeling income. These explanatory variables are, among others, gender (`gender`), the equivalized household size (`eqsize`) as well as financial resources like the employees cash (`cash`) or unemployment benefits (`unempl_ben`). Table 4.2 gives an overview of possible model covariates.

The sample data set `eusilcA_smp` is a household sample from the `eusilcA_pop` population that includes 1945 observations. The sample is drawn by stratified random sampling where the districts define the strata. For the 75% largest districts (in terms of number of households) 10% of the households were selected and the maximum number of sampled households is equal to 200 in any given district. Consequently, the 25% smallest districts do not have any observation in the sample. Summaries of state and district-specific sample sizes are given below.

```
R> data("eusilcA_smp")
R> table(eusilcA_smp$state)
```

Burgenland	Carinthia	Lower Austria	Salzburg
31	162	387	163
Styria	Tyrol	Upper Austria	Vienna
337	173	392	200
			Vorarlberg
			100

Variable	Meaning	Scale level
Target variable		
eqIncome	The equalized household income.	Numeric
Domain identifiers		
state	Austrian states.	Factor
district	Austrian districts.	Factor
Explanatory variables		
eqsize	The equalized household size according to the modified OECD scale.	Numeric
gender	The person's gender (levels: <i>female</i> and <i>male</i>).	Factor
cash	Employee cash or near cash income.	Numeric
self_empl	Cash benefits or losses from self-employment (net).	Numeric
unempl_ben	Unemployment benefits (net).	Numeric
age_ben	Old-age benefits (net).	Numeric
surv_ben	Survivor's benefits (net).	Numeric
sick_ben	Sickness benefits (net).	Numeric
dis_ben	Disability benefits (net).	Numeric
rent	Income from rental of a property or land (net).	Numeric
fam_allow	Family/children related allowances (net).	Numeric
house_allow	Housing allowances (net).	Numeric
cap_inv	Interest, dividends, profit from capital investments in unincorporated business (net).	Numeric
tax_adj	Repayments/receipts for tax adjustment (net).	Numeric
Design variable		
weight	Sampling weight.	Numeric

Table 4.2: Variables of the two data sets in package **emdi**. Note that the population data set does not contain a variable for the sampling weights.

```
R> summary(as.numeric(table(eusilcA_smp$district)))
```

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
14.00 17.00 22.50 27.79 29.00 200.00
```

District-specific sample sizes (in contrast to state-specific) are quite small with 25% of districts having no sample data at all. Hence, the use of SAE methods may be useful in this case. In Section 4.4 we discuss the estimation of regional indicators based on these data sets in detail.

In addition to SAE methods, package **emdi** provides a function called `map_plot` that produces maps of the estimated indicators. In order to demonstrate the use of the function `map_plot` package **emdi** contains a shape file for the 94 Austrian districts which is downloaded from the SynerGIS website (Bundesamt für Eich- und Vermessungswesen, 2017). This shape file is saved in `.rda` format and the object `shape_austria_dis` is a `SpatialPolygonsDataFrame`. For more information about this class we refer to Bivand et al. (2013).

4.4 Basic design and core functionality

Section 4.2 presented the statistical methodology that uses either direct estimation or the model-based EBP approach. In package **emdi** direct and model-based estimation are implemented

Position	Name	Short description	Available for	
			direct	model
1	<code>ind</code>	Point estimates for indicators per domain	✓	✓
2	<code>MSE</code>	Variance/MSE estimates per domain	✓	✓
3	<code>transform_param</code>	Transformation and shift parameters		✓
4	<code>model</code>	Fitted linear mixed-effects model as <code>lme</code> object		✓
5	<code>framework</code>	List with 8 components describing the data	✓	✓
6	<code>transformation</code>	Type of transformation		✓
7	<code>method</code>	Estimation method for transformation parameter		✓
8	<code>fixed</code>	Formula of fixed effects used in the nested error linear regression model		✓
9	<code>call</code>	Image of the function call that produced the object	✓	✓
10	<code>successful_bootstraps</code>	A matrix with domains as rows, indicators as columns and the number of corresponding successful bootstraps	✓	

Table 4.3: Components of `emdi` objects. All explanations can be found in the documentation of the `emdi` object in the package.

with functions `direct` and `ebp`, respectively. A key benefit offered by **emdi** is the flexibility for producing, assessing, presenting and exploring the estimates. This is achieved by using the following commands:

1. Estimate domain indicators including MSE estimation: `direct` and `ebp`
2. Get summary statistics and model diagnostics: `summary` and `plot`
3. Extract and compare the indicators of interest: `estimators` and `compare`
4. Visualize the estimated indicators: `map_plot`
5. Export the results to ExcelTM: `write.excel`

The package **emdi** uses the S3 object system (Chambers and Hastie, 1992). All objects created in the package **emdi** by an estimation function (`direct` and `ebp`) share the class `emdi`. Objects of class `emdi` comprise ten components, which are presented in Table 4.3. Some of these components are specific only to one of the estimation methods, such that they are `NULL` for the other one. These components are indicated in the second column of Table 4.3. Depending on the estimation method, the `emdi` object is also of class `direct` or `model`. Thus, the commands can be tailored to the estimation method, e.g., model diagnostics (provided by the command `plot`) are only suitable when a model-based approach is used. In what follows the estimation functions are presented and **emdi** functionalities are illustrated.

Arguments	Short description	Default
<code>y</code>	Target variable	
<code>smp_data</code>	Survey data	
<code>smp_domains</code>	Domain identifier	
<code>weights</code>	Sampling weights	No weights
<code>design</code>	Variable indicating strata	No design
<code>threshold</code>	Threshold for poverty indicators	60% of the median of the target variable
<code>var</code>	Variance estimation	No variance estimation
<code>boot_type</code>	Type of bootstrap: naive or calibrate	Naive
<code>B</code>	Number of bootstrap populations	50
<code>seed</code>	Seed for random number generator	123
<code>X_calib</code>	Calibration variables	None
<code>totals</code>	Population totals	None
<code>custom_indicator</code>	Customized indicators	None
<code>na.rm</code>	Deletion of observations with missing values	No deletion

Table 4.4: Input arguments for function `direct`. All explanations can also be found in the documentation of the `direct` function in the package.

4.4.1 Estimation of domain indicators

As far as possible, the two estimation functions (`direct` and `ebp`) have the same structure and variable names, which helps to simplify their use. For function `direct`, the user has to specify three arguments (see Table 4.4), that include the target variable, the sample data set, and the variable name that defines the domain identifier in the sample data. For the remaining arguments suitable defaults are defined. The EBP approach is implemented in **emdi**, using function `ebp`. As shown in Table 4.5, the user has to specify five arguments that include the structure of the fixed effects of the nested error linear regression model, the two data sets (population and sample), and the variable names that define the domain identifiers in each data set. For the remaining arguments suitable defaults are defined. Following Molina and Rao (2010), the number of Monte Carlo iterations L and the number of bootstrap populations B are set to 50 by default. In practice, we recommend using larger values for example, $L \geq 200$ and $B \geq 200$. The choice of a transformation is simplified since the user only has to choose the type of transformation. The shift parameter s and the optimal transformation parameter λ in the case of using the Box-Cox transformation are automatically estimated. This distinguishes **emdi** from package **sae** (Molina and Marhuenda, 2015) where the user has to select the transformation parameters manually. Since the Box-Cox transformation includes the no transformation and logarithmic transformation as special cases, this family of transformations is chosen as the default option.

Example using Austrian districts:

For illustrating the functions of package **emdi** we estimate indicators using the data sets described in Section 4.3. The target variable is the equivalized income (`eqIncome`) and the regional level of interest are Austrian districts included in variable `district`. For direct estimation of the indicators the user has to specify these two arguments and the sample data

Arguments	Short description	Default
<code>fixed</code>	Fixed effects formula of the nested error regression model	
<code>pop_data</code>	Census or administrative data	
<code>pop_domains</code>	Domain identifier for population data, <code>pop_data</code>	
<code>smp_data</code>	Survey data	
<code>smp_domains</code>	Domain identifier for sample data, <code>smp_data</code>	
<code>L</code>	Number of Monte Carlo iterations	50
<code>threshold</code>	Threshold for poverty indicators	60% of the median of the target variable
<code>transformation</code>	Type of transformation: no, log or Box-Cox	Box-Cox
<code>interval</code>	Interval for the estimation of the optimal transformation parameter	(-1,2)
<code>MSE</code>	Mean Squared Error (MSE) estimation	No MSE estimation
<code>B</code>	Number of bootstrap populations	50
<code>seed</code>	Seed for random number generator	123
<code>boot_type</code>	Type of bootstrap: parametric or wild	Parametric
<code>parallel_mode</code>	Mode of parallelization	Automatic
<code>cpus</code>	Number of kernels for parallelization	1
<code>custom_indicator</code>	Customized indicators	None
<code>na.rm</code>	Deletion of observations with missing values	No deletion

Table 4.5: Input arguments for function `ebp`. All explanations can also be found in the documentation of the `ebp` function in the package.

set called `eusilcA_smp`. In addition, several other arguments are defined as shown below. We account for the sampling design by including the sampling weights in the estimation. Furthermore, we set the `threshold` argument to 60% of the median of equivalized income that – in this example – equals 10885.33 and we are also interested in obtaining the variance estimates of the indicators.

```
R> emdi_direct <- direct(y = "eqIncome",
+   smp_data = eusilcA_smp, smp_domains = "district",
+   weights = "weight", threshold = 10885.33, var = TRUE)
```

The R object `emdi_direct` is of classes `emdi` and `direct`.

An example of using the `ebp` method for computing point and MSE estimates for the predefined indicators and two custom indicators, namely the minimum and maximum equivalized income is provided below:

```
R> emdi_model <- ebp(fixed = eqIncome ~ gender + eqsize +
+   cash + self_empl + unempl_ben + age_ben + surv_ben +
+   sick_ben + dis_ben + rent + fam_allow + house_allow +
+   cap_inv + tax_adj,
+   pop_data = eusilcA_pop, pop_domains = "district",
+   smp_data = eusilcA_smp, smp_domains = "district",
+   threshold = 10885.33, MSE = TRUE,
```

```
+   custom_indicator =
+       list(my_max = function(y, threshold) {max(y)},
+            my_min = function(y, threshold) {min(y)}))
```

In contrast to the direct estimation, the user also has to choose the auxiliary variables to be included in the nested error linear regression model. The variables that are chosen to explain the equivalized income are demographics as gender and the equivalized household size but also financial benefits and allowances as for example cash income, unemployment benefits and capital investment. Furthermore, model-based estimation requires the use of both, population (`eusilcA_pop`) and sample (`eusilcA_smp`) data and the domain identifiers. For enabling the comparison between direct and model-based estimates of the indicators of interest we use the same threshold as in the direct estimation. MSE estimates are returned by setting the `MSE` argument to `TRUE`. The final R object `emdi_model` is of classes `emdi` and `model`. For this object we show in the following sections the **emdi** functionalities.

4.4.2 Summary statistics and model diagnostics

R-users typically use a `summary` method for summarizing the results. For `emdi` objects the summary outputs differ depending on the two classes. The summary for objects obtained by direct estimation gives information about the number of domains in the sample, the total and domain-specific sample sizes. The summary for model-based objects is more extensive. In addition to information about the sample sizes, information about the population size and the number of out-of-sample domains is provided. Since model-based SAE relies on prediction under the model, including model diagnostics in **emdi** is important for users. A first measure to consider when evaluating the working model is the well known R^2 . Nakagawa and Schielzeth (2013) provide a generalization of this measure for linear mixed models. A marginal R^2 and a conditional (a measure that accounts for the random effect) R^2 are implemented via function `r.squaredGLMM` in package **MuMIn** (Barton, 2018). The `summary` method uses this function to calculate and present both measures. For the EBP and model-based SAE methods in general the validity of parametric assumptions is crucial. Therefore, **emdi** also outputs residual diagnostics. In particular, results include the skewness and kurtosis of both sets of residuals (random effects and unit-level) and the results from using the Shapiro-Wilk test for normality (test statistic and p-value). The intra-cluster correlation (ICC) coefficient is further used for assessing the remaining unobserved heterogeneity. Finally, the `summary` command gives information about the selected transformation. If the user opts for a Box-Cox transformation, the transformation parameter λ and the shift parameter s are reported.

In addition to the diagnostics provided by `summary`, **emdi** enables the use of graphical diagnostics (see Figure 4.1). The `plot` method outputs graphics of residual diagnostics. The first set of plots (Figure 4.1a) are normal quantile-quantile (Q-Q) plots of Pearson unit-level residuals and standardized random effects. Figure 4.1b and 4.1c are kernel density plots of the distribution of the two sets of residuals contrasted against a standard normal distribution. Outliers can have a significant impact on the model fit and hence on prediction. Hence, a Cook's distance plot is also available (Figure 4.1d), in which the three largest values of the standardized residuals are identified alongside the case identification number for further investigation.

Finally, if a Box-Cox transformation is used, a plot of the profile log-likelihood that shows the value of the transformation parameter for which the likelihood is maximized is also produced (Figure 4.1e). The user can customize the format of all plots. Method `plot` accepts the parameter `label` with the predefined values `blank` (deletes all labels) and `no_title` (axis labels are given, but no plot titles). In addition, a user-defined list that contains specific labels for each plot list can be given. Another parameter available is `color` which accepts a vector with two color specifications. The first color defines the lines in Figure 4.1a, 4.1d and 4.1e and the second one specifies the color of the shapes in Figure 4.1b and 4.1c. For the likelihood plot the range in which the likelihood should be computed can be specified by using the parameter `range`. The appearance of the plots benefits from the use of the **ggplot2** package (Wickham, 2009). Hence, `plot` accepts a `gg_theme` argument that allows for all customization options of `theme` that is a tool for modifying non-data components of a plot.

Example using Austrian districts:

In order to check the diagnostics in our example we use the `summary` and the `plot` methods. The summary output of the object `emdi_model` is presented below.

```
R> summary(emdi_model)
```

```
Empirical Best Prediction
```

```
Call:
```

```
ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
unempl_ben + age_ben + surv_ben + sick_ben + dis_ben + rent +
fam_allow + house_allow + cap_inv + tax_adj,
pop_data = eusilcA_pop, pop_domains = "district",
smp_data = eusilcA_smp, smp_domains = "district",
threshold = 10885.33, MSE = TRUE,
custom_indicator =
  list(my_max = function(y, threshold) {max(y)},
       my_min = function(y, threshold) {min(y)}))
```

```
Out-of-sample domains: 24
```

```
In-sample domains: 70
```

```
Sample sizes:
```

```
Units in sample: 1945
```

```
Units in population: 25000
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample_domains	14	17.0	22.5	27.78571	29.00	200
Population_domains	5	126.5	181.5	265.95745	265.75	5857

```

Explanatory measures:
Marginal_R2 Conditional_R2
  0.6325942      0.709266

Residual diagnostics:
                Skewness Kurtosis Shapiro_W      Shapiro_p
Error          0.7523871  9.646993 0.9619824 3.492626e-22
Random_effect 0.4655324  2.837176 0.9760574 1.995328e-01

ICC:  0.2086841

Transformation:
Transformation Method Optimal_lambda Shift_parameter
      box.cox    reml      0.6046901              0

```

This output helps to justify the use of a model-based approach for SAE in this specific example. On the one hand, 24 out of 94 districts are out-of-sample such that direct estimates cannot be produced for these districts. Furthermore, the sample sizes in the districts are rather small with a median of 22.5 households and vary between a minimum of 14 households and a maximum of 200 households. The explanatory power of the selected covariates is high with the conditional R^2 , the measure that jointly considers the fixed and the random effect, of around 71%. The ICC of 20.9% further justifies the inclusion of a random effect. The normality tests show that normality is rejected for the unit-level error term but not for the random effect. The use of transformations helps to reduce the skewness of the distribution of the error terms. The optimal transformation parameter is 0.6 indicating that neither using the untransformed income or the logarithmic transformation of income would be appropriate for this data set. The plots in Figure 4.1 used for residual analyses of the object `emdi_model` can be produced as follows,

```

R> plot(emdi_model, label = "no_title",
+       color = c("red3", "red4"))

```

The Q-Q plots and the densities of the two error terms confirm that normality seems to be reasonable for the random effect but not for the unit-level error term. Furthermore, the Cook's distance plot identifies possible outliers. The last plot shows the optimal transformation parameter, which is the maximum of the profile log-likelihood.

4.4.3 Selection and comparison of indicators

Package `emdi` returns a set of predefined and customized indicators. The ten predefined indicators are summarized in Table 4.1. However, the user may only be interested in some of these or only in individually defined (customized) indicators. A function called `estimators` helps the user to select the indicator or indicators of interest. This is done by using the `indicator` argument that takes a vector of indicator names as an argument, but in addition also accepts keywords defining predefined groups; for example, the keyword `custom` returns only user-defined indicators. In addition to variance and MSE estimates, NSI often use the CV as an

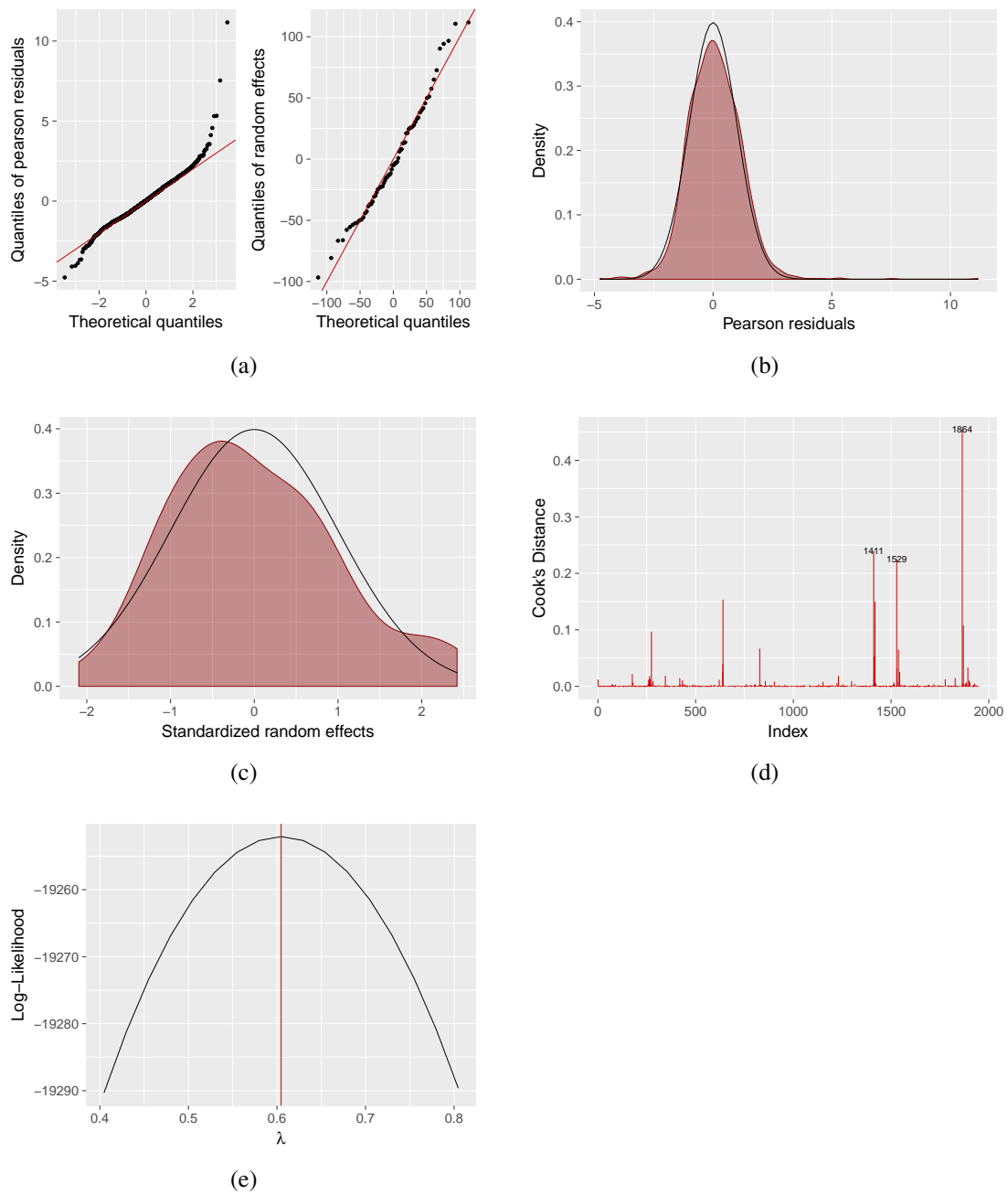


Figure 4.1: Graphics obtained by using `plot(emdi_model)`. (a) shows normal Q-Q plots of the unit-level errors and the random effects. (b) and (c) show kernel density estimates of the distributions of standardized unit-level errors and standardized random effects compared to a standard normal distribution (black density). The Cook's distance plot is displayed in (d) whereby the index of outliers is labeled. The profile log-likelihood for the optimal parameter value of the Box-Cox transformation is shown in (e).

additional measure of the quality of the estimates. Estimated CVs as defined in Section 4.2 can be returned alongside MSE estimates.

It is often important to compare model-based and direct estimates. Direct estimates do not depend on the use of a model and hence the analyst should be interested in deriving model-based estimates that are close to direct estimates. Comparing model-based to direct estimates offers an internal validation procedure for checking whether the use of a model leads to unreasonable estimates. Package **emdi** provides a function called `compare_plot` that returns two plots, a scatter plot according to Brown et al. (2001) and a line plot. The scatter plot shows the direct and model-based point estimates, the fitted regression line, and the identity line. The closer the regression line is to the identity line, the closer the estimates are. The line plot is shown for domains ordered by the sample size. Thus, the user can see how the model-based estimates track the direct estimates across domains. In accordance with the function `estimators` the user can choose which indicators are compared by using the `indicator` argument. Similarly to the diagnostic plots, the user can modify the layout of the two plots. The label options are also `blank` (deletes all labels) and `no_title` (axis labels are given, but no plot titles). The color, the shape of the points and the type of the lines can be changed by using arguments `color`, `shape` and `line_type`, respectively.

Example using Austrian districts:

We illustrate how to estimate the median of equivalized income and the Gini coefficient and the corresponding CV estimates for the first 6 districts in Austria.

```
R> head(estimators(emdi_model, indicator = c("Gini", "Median"),
+ MSE = FALSE, CV = TRUE))
```

	Domain	Gini	Gini_CV	Median	Median_CV
1	Eisenstadt-Umgebung	0.2214688	0.09790984	25414.07	0.10381883
2	Eisenstadt (Stadt)	0.2872751	0.06110093	49274.84	0.07673551
3	Güssing	0.1906263	0.13046770	16718.13	0.12732081
4	Jennersdorf	0.2098103	0.15371048	12869.55	0.17815504
5	Mattersburg	0.2091353	0.10851693	20102.09	0.12764578
6	Neusiedl am See	0.1865026	0.05934130	18386.83	0.06346778

For these districts, the Gini coefficient and the median income are highest in Eisenstadt (Stadt). The lowest Gini is in Neusiedl am See and the lowest median in Jennersdorf. Furthermore, it can be noted that none of the CVs is above 20%. This threshold is used by the ONS in UK in order to decide if estimates can be reported.

The plots in Figure 4.2 are obtained by

```
R> compare_plot(emdi_direct, emdi_model,
+ indicator = c("Gini", "Median"), label = "no_title",
+ color = c("red3", "blue"))
```

The scatter plots highlight that the disparity of the fitted regression line from the identity line is higher for the Gini coefficient than for the median. The model-based estimates do not track

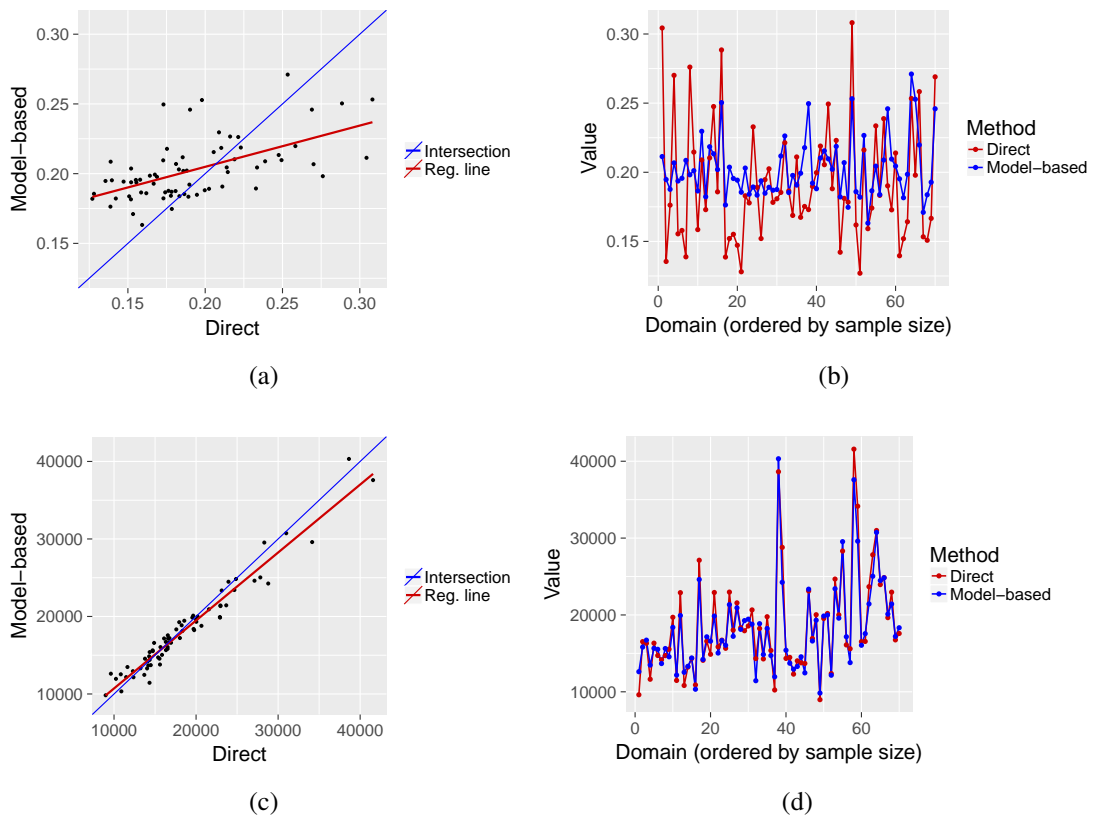


Figure 4.2: Graphics obtained by using `compare_plot(emdi_model)`. (a) and (c) show the scatter plots of the direct and model-based estimates for the Gini coefficient (top) and the median (bottom), respectively. (b) and (d) are line plots of the same estimates where the domains are ordered by increasing sample size.

pop_data_id	shape_id
ID of domain 1 in the emdi obj	ID of domain 1 in the shape file
ID of domain 2 in the emdi obj	ID of domain 2 in the shape file
ID of domain 3 in the emdi obj	ID of domain 3 in the shape file
⋮	⋮

Table 4.6: Example of a mapping table for argument `map_tab` in function `map_plot` in **emdi**.

the direct estimates and show also a lower variability across the domains. In contrast, the direct and model-based estimates for the median are close to each other. Especially for large domains the difference is negligible.

4.4.4 Mapping of the estimates

In SAE maps are a natural way to present the estimates as they help describing the spatial distribution of issues like poverty and inequality. Creating maps can be demanding or laborious in practice. Package **emdi** includes function `map_plot` that simplifies the creation of maps. Given a spatial polygon provided by a shape file and a corresponding `emdi` object `map_plot` produces maps of selected indicators and corresponding MSE and CV estimates. The parameters `MSE`, `CV` and `indicator` correspond to those in the `estimators` function. As Wickham (2009) points out the matching of domain identifiers in the statistical data to the corresponding identifiers in the spatial data (shape file) is challenging and general solutions are hard to obtain. The function `map_plot` in **emdi** allows for an argument `map_tab` when the identifiers do not match. The user must define a mapping table (cf. Table 4.6) for the argument `map_tab` in the form of a data frame that matches the domain variable in the population data set with the domain variable in the shape file. If the domain identifiers in both data sources match, this table is not required. The handling of the spatial shape files can be done using package **maptools** (Bivand and Lewin-Koh, 2017) in combination with package **rgeos** (Bivand and Rundel, 2017). Alternative approaches are provided by the packages **rgdal** (Bivand et al., 2018) and **sf** (Pebesma, 2018). For general information on how to work with spatial data and shape files we refer the reader to Bivand et al. (2013).

Example using Austrian districts:

The steps for obtaining a map of median income in Austrian districts and the corresponding CVs are outlined below. First, the shape file needs to be loaded.

```
R> load_shapeaustria()
```

Then, two maps are created (cf. Figure 4.3).

```
R> map_plot(emdi_model, MSE = FALSE, CV = TRUE,
+   map_obj = shape_austria_dis, indicator = "Median",
+   map_dom_id = "PB")
```

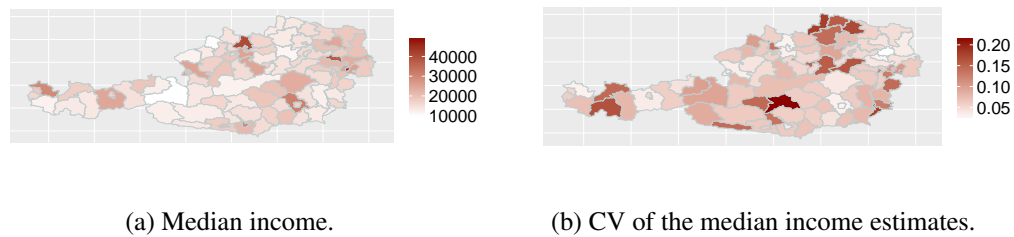


Figure 4.3: Maps of point estimates and CVs of the median income for 94 districts in Austria.

As the domain identifiers in the data set and shape file already match, the argument `map_tab` is not required. For an example where the argument `map_tab` needs to be specified, we refer the reader to `help(map_plot)`.

The map of the median equivalized income in Figure 4.3 indicates differences across Austrian districts. The richest district appears to be Eisenstadt (Stadt) followed by Urfahr-Umgebung. Furthermore, throughout the country some districts have a relatively low median income like Zell am See and Schärding. The map of the CVs shows that most districts have a CV below 20%. The highest CVs occur in the out-of-sample domains.

4.4.5 Exporting the results

Exporting the results from R to other widely used software such as Excel™ is important for users. For doing so a large set of well established tools already exists. Nevertheless, exporting all model information, including the information contained in the summary output is not straightforward. Function `write.excel` creates a new Excel™ file that contains the summary output in the first sheet and the results from the selected estimators in the following sheet. Again the parameters `MSE`, `CV` and `indicator` correspond to those in the `estimators` function. The link with the Excel™ file format is done by using the package `openxlsx` (Walker, 2017). This package does not require a Java™ installation, which offers an advantage over the use of the `xlsx` package (Dragulescu, 2014) because Java™ may be seen as a potential security threat. Nevertheless, package `openxlsx` (Walker, 2017) needs a zipping application available to R. Under Microsoft Windows™ this can be achieved by installing RTools while under macOS™ or Linux™ such an application is available by default. In addition to exporting the results to Excel™, `emdi` also provides an option to export output directly as OpenDocument Spreadsheets via the function `write.ods`.

Example using Austrian districts:

Excel™ outputs of model-based estimates for Austrian districts can be obtained by the following command.

```
R> write.excel(emdi_model, file = "excel_output.xlsx",
+   indicator = "Median", MSE = FALSE, CV = TRUE)
```

The output is presented in Figure 4.4 and shows that also the Excel™ user receives the same diagnostics from the summary and results for selected estimates. The summary output is described in detail in Section 4.4.2.

row_names	Count
out of sample domains	24
in sample domains	70
out of sample observations	25000
in sample observations	1945

row_names	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample_domains	14	17	22.5	27.7857143	29	200
Population_domains	5	126.5	181.5	265.957447	265.75	5857

Transformation	Method	Optimal_lambda	Shift_parameter
box.cox	reml	0.604690114	0

row_names	Skewness	Kurtosis	Shapiro_W	Shapiro_p
Error	0.75238713	9.646993319	0.96198243	3.4926E-22
Random_effect	0.46553236	2.837176266	0.976057372	0.1995328

	Marginal_R2	Conditional_R2
	0.63259425	0.709265978

	A	B	C
1	Domain	Median	Median_CV
2	Eisenstadt-Umgebung	25414.0656	0.10381883
3	Eisenstadt (Stadt)	49274.8446	0.07673551
4	Güssing	16718.1284	0.12732081
5	Jennersdorf	12869.5499	0.17815504
6	Mattersburg	20102.0868	0.12764578
7	Neusiedl am See	18386.8329	0.06346778

Figure 4.4: Export of the summary output and estimates to Excel™.

4.5 Additional features

In addition to those features that are essential for estimating regional indicators, package **emdi** offers to incorporate external indicators and increases the computational efficiency of the MSE estimation by parallel computing. In this section we show how users can bring indicators from other R packages into **emdi** and how parallel computing can help with reducing the computational burden.

4.5.1 Incorporating an external indicator

A feature we should pay attention to is the ease by which indicators of other R packages can be brought into **emdi**. This is demonstrated by using the Theil index from the R package **ineq** (Zeileis, 2014). The Theil index describes economic inequality and thus can be also used in the application with the data of this paper. It belongs to a family of generalized entropy inequality measures and can be expressed by

$$\text{Theil}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{y_{ij}}{\bar{y}} \log \left(\frac{y_{ij}}{\bar{y}} \right),$$

where $\bar{y} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ (Cowell, 2011). The Theil index takes values from 0 to ∞ with 0 indicating equality and higher values increasing inequality (World Bank Institute, 2005). As the function **ineq** only requires a numeric vector of the target variable, it can be straightforwardly wrapped into a form usable within the **direct** or **ebp** functions. Using the function **direct** the Theil index can be estimated as follows.

First, the package **ineq** needs to be installed and loaded.

```
R> install.packages("ineq")
R> library("ineq")
```

Subsequently, the function **ineq** with `type = "Theil"` can be given to the argument `custom_indicator`. As the function **direct** needs the arguments `y`, `weights` and `threshold`, these arguments have to be also specified in the newly defined function.

```
R> my_theil <- function(y, weights, threshold) {
+   ineq(x = y, type = "Theil")
+ }
```

The argument `custom_indicator` needs to include a named list of self-defined indicators.

```
R> my_indicators <- list(theil = my_theil)
R> emdi_direct2 <- direct(y = "eqIncome",
+   smp_data = eusilcA_smp, smp_domains = "district",
+   weights = "weight", var = TRUE,
+   custom_indicator = my_indicators)
```

As the Theil index is now part of the `emdi` object, all methods shown in Section 4.4 can be also used for this newly defined inequality indicator. For instance, by estimating a customized indicator via function `direct` a bootstrap variance estimator is used and the `subset` method can be applied in order to get results for certain districts.

```
R> select_theil <- estimators(emdi_direct2,
+   indicator = "theil", CV = TRUE)
R> subset(select_theil, Domain == "Wien")
```

```
      Domain      theil  theil_CV
67  Wien 0.1202542 0.1108617
```

4.5.2 Parallelization

Bootstrapping the MSE can be very costly in terms of computation time and the possibilities of speeding up are limited when staying within R. Nevertheless, as the bootstrap procedures described in Section 4.2.2 and Appendix C.1 consist of B independent iterations, they are suitable for efficient parallel computing. In this particular case, parallelization may be described as follows:

1. The user predefines how many parallel processes (`cpus`) and bootstrap iterations (B) should be used in function `ebp`.
2. The bootstrap iterations are equally distributed on the parallel processes.
3. In each process the differences between EBP point estimates and the pseudo true values $\widehat{\Delta I}_{i,b} = \hat{I}_{i,b}^{EBP} - I_{i,b}$ (compare e.g., Appendix C.1) are calculated. This is done on different central processing units (CPU) at the same time (parallel computing).
4. The results $\widehat{\Delta I}_{i,b}$ from all processes are combined and the MSE is estimated by
$$\widehat{MSE}(\hat{I}_i^{EBP}) = B^{-1} \sum_{b=1}^B (\widehat{\Delta I}_{i,b})^2.$$

In R there are numerous ways and packages for implementing parallel computing. The most used package in this context is **parallel** (R Core Team, 2018), which mainly builds on the work of packages **snow** (Tierney et al., 2016) and **multicore** (Urbanek, 2014). These packages follow two different approaches for parallelization. Package **snow** launches a new version of

R on each core. Those versions communicate with the master process through the so-called “socket”. Therefore, we will proceed calling this way of parallelization the socket approach. The second approach is called “forking” and is the approach developed in the **multicore** package. Forking duplicates the entire current version of R and shifts it to a new core. Forking has one crucial advantage: all slave processes share the same memory with the master process for any object that is not modified. This feature makes it very fast. Its disadvantage is that it is not available on Microsoft Windows™ operating systems. The **parallel** package allows for both approaches but uses different functions. These functions are given an unified interface by the package **parallelMap** (Bischl and Lang, 2015). This interface for parallelization is used in **emdi**. In the `ebp` function the parallelization approach defaults to socket if a Microsoft Windows™ OS is detected and to forking otherwise. The parallelization is activated by setting the `cpus` argument to an integer value larger than 1. In the example below the computation time is measured when the number of CPU is set equal to 1 and to 2, respectively:

```
R> system.time(emdi_modell <- ebp(fixed = eqIncome ~ gender +
+   eqsize + cash + self_empl + unempl_ben + age_ben +
+   surv_ben + sick_ben + dis_ben + rent + fam_allow +
+   house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
+   pop_domains = "district", smp_data = eusilcA_smp,
+   smp_domains = "district", threshold = 10885.33,
+   MSE = TRUE, seed = 100, cpus = 1))
```

user	system	elapsed
155.86	0.09	157.36

```
R> system.time(emdi_model2 <- ebp(fixed = eqIncome ~ gender +
+   eqsize + cash + self_empl + unempl_ben + age_ben +
+   surv_ben + sick_ben + dis_ben + rent + fam_allow +
+   house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
+   pop_domains = "district", smp_data = eusilcA_smp,
+   smp_domains = "district", threshold = 10885.33,
+   MSE = TRUE, seed = 100, cpus = 2))
```

user	system	elapsed
3.62	0.45	89.45

The return value `elapsed` from function `system.time` informs the user about the real time that has passed from submitting the command until completion. Hence, the time comparison shows that two parallel processes reduce the time that is needed for the `ebp` function to run approximately by half. Please note that computation times are not replicable.

Despite the advantages in terms of computation time, parallelization comes with a major drawback. The reproducibility of results that depends on random number generations is non trivial. The usual `set.seed()` command that is used in R to ensure reproducibility is not sufficient due to the different R sessions used in parallel computing. In the socket approach, the

function `clusterSetRNGStream()` from the **parallel** package is used to provide reproducible random number streams to each process that are far apart from each other. Therefore, all processes would produce different but reproducible random numbers. When using the forking approach, reproducibility can be more easily achieved by simply using a different random number generator. In the `ebp` function, `set.seed(seed, kind = "L'Ecuyer")` is used to set the random number generation to L'Ecuyer (L'Ecuyer et al., 2002) which is based on L'Ecuyer (1999). The multiple substreams of random numbers are created by the **rstream** package (Leydold, 2017) in both approaches. Please note that results obtained from parallel computation are only reproducible if the same number of processes and the same parallelization approach are used. The reproducibility is demonstrated below by reproducing the results with `cpus` equal to 2.

```
R> emdi_model22 <- ebp(fixed = eqIncome ~ gender + eqsize +
+   cash + self_empl + unempl_ben + age_ben + surv_ben +
+   sick_ben + dis_ben + rent + fam_allow + house_allow +
+   cap_inv + tax_adj, pop_data = eusilcA_pop,
+   pop_domains = "district", smp_data = eusilcA_smp,
+   smp_domains = "district", threshold = 10885.33,
+   MSE = TRUE, seed = 100, cpus = 2)
```

```
R> all.equal(emdi_model2, emdi_model22)
```

```
[1] TRUE
```

4.6 Conclusion and future developments

In this paper we show how the **emdi** package can simplify the application of SAE methods. This package is, to the best of our knowledge, the first R SAE package that supports the user beyond estimation in the production of complex, non-linear indicators. Another important feature is that data-driven transformation parameters are estimated automatically. Estimating the uncertainty of small area estimates is achieved by using both parametric bootstrap and semi-parametric wild bootstrap. The additional uncertainty due to the estimation of the transformation parameter is also captured in MSE estimation. Customized parallel computing is included for reducing the computational time. The complexity in applying SAE methods is considerably reduced, useful diagnostic tools are incorporated and the user is also supported by the availability of tools for presenting, visualizing and further processing the results. For instance, the model summary and results can be exported to Excel™ and to OpenDocument Spreadsheets. Since **emdi** makes the application of SAE methods in R almost as simple as fitting a linear or a generalized linear regression model, it also has the potential to close the gap between theoretical advances in SAE and their application by practitioners.

Additional features will be integrated in future versions of the package. Firstly, the implementation of alternative SAE methods will increase the usage of the package. For example, the World Bank (Elbers et al., 2003) and M-Quantile (Chambers and Tzavidis, 2006; Tzavidis et al.,

2010) methods complement the EBP approach (Molina and Rao, 2010) for estimating disaggregated complex, non-linear indicators. Secondly, including additional evaluation and diagnostic tools for comparing direct and model-based estimates will assist the user with deciding which estimation method should be preferred. Thirdly, currently **emdi** includes only some possible types of transformations and one estimation method for the transformation parameter, namely REML. Future versions of the package will include a wider range of transformations (e.g., log shift and dual power transformations) and alternative estimation methods (minimization of the skewness or measures of symmetry) for the transformation parameter.

Acknowledgments

Rojas-Perilla, Schmid and Tzavidis gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council. The work of Kreutzmann and Schmid has been also supported by the German Research Foundation within the project QUESSAMI (281573942) and by the MIUR-DAAD Joint Mobility Program (57265468). The numerical results are not official estimates and are only produced for illustrating the methods. We thank the editor and the referees for their constructive comments that helped to improve the paper.

Appendix C

C.1 Semi-parametric wild bootstrap

The semi-parametric wild bootstrap is implemented as follows,

1. Fit model 4.1 (using an appropriate transformation for y_{ij}) to obtain estimates $\hat{\beta}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$.
2. Calculate the sample residuals by $\hat{\epsilon}_{ij} = y_{ij} - \mathbf{x}_{ij}^\top \hat{\beta} - \hat{u}_i$.
3. Scale and center these residuals using $\hat{\sigma}_e$. The scaled and centered residuals are denoted by $\hat{\epsilon}_{ij}$.
4. For $b = 1, \dots, B$

- (a) Generate $u_i^{(b)} \stackrel{iid}{\sim} N(0, \hat{\sigma}_u^2)$.
- (b) Calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^\top \hat{\beta} + u_i^{(b)}$.
- (c) Match $\eta_{ij}^{(b)}$ with the set of estimated linear predictors $\{\hat{\eta}_k | k \in n\}$ from the sample by using

$$\min_{k \in n} \left| \eta_{ij}^{(b)} - \hat{\eta}_k \right|$$

and define \tilde{k} as the corresponding index.

- (d) Generate weights w from a distribution satisfying the conditions in Feng et al. (2011) where w is a simple two-point mass distribution with probabilities 0.5 at $w = 1$ and $w = -1$, respectively.
- (e) Calculate the bootstrap population as $T(y_{ij}^{(b)}) = \mathbf{x}_{ij}^\top \hat{\beta} + u_i^{(b)} + w_{\tilde{k}} |\hat{\epsilon}_{\tilde{k}}^{(b)}|$.
- (f) Back-transform $T(y_{ij}^{(b)})$ to the original scale and compute the bootstrap population value $I_{i,b}$.
- (g) Select the bootstrap sample and use the EBP method as described above.
- (h) Obtain $\hat{I}_{i,b}^{EBP}$.

$$5. \widehat{MSE}_{Wild} \left(\hat{I}_i^{EBP} \right) = B^{-1} \sum_{b=1}^B \left(\hat{I}_{i,b}^{EBP} - I_{i,b} \right)^2.$$

A simulation study assessing the performance of the semi-parametric wild bootstrap is presented in Rojas-Perilla et al. (2017).

C.2 Reproducibility

The results presented in this paper were obtained under R version 3.4.4 on a 64-bit platform under Microsoft Windows 7™. The installed packages are listed in Table C.1. A snapshot of the corresponding repository was created with the package **packrat** (Ushey et al., 2018) and is available from the authors' GitHub folder (<https://github.com/SoerenPannier/emdi.git>). To make use of this repository Git must be installed. The authors recommend the following workflow:

- Use the new project functionality from RStudio.
- Choose checkout from version control and select Git.
- Enter the repository URL: `https://github.com/SoerenPannier/emdi.git`.
- Wait until **packrat** finishes the initialization process.
- Restart RStudio.
- Enter the R command `packrat::restore()`.
- After the package installation has finished all packages are installed as documented in Table C.1.

Package	Version	Package	Version	Package	Version
assertthat	0.2.0	mgcv	1.8-23	stringi	1.1.7
backports	1.1.2	mime	0.5	stringr	1.3.0
BBmisc	1.11	minqa	1.2.4	testthat	2.0.0
BH	1.66.0-1	moments	0.14	tibble	1.4.2
boot	1.3-20	MuMIn	1.40.4	utf8	1.1.3
brew	1.0-6	munsell	0.4.3	viridisLite	0.3.0
cellranger	1.1.0	nlme	3.1-131.1	whisker	0.3-2
checkmate	1.8.5	nloptr	1.0.4	withr	2.1.2
cli	1.0.0	openssl	1.0.1	xml2	1.2.0
colorspace	1.3-2	openxlsx	4.0.17	base	3.4.4
commonmark	1.4	packrat	0.4.9-1	boot	1.3-20
crayon	1.3.4	parallelMap	1.3	class	7.3-14
curl	3.1	pillar	1.2.1	cluster	2.0.6
desc	1.1.1	pkgconfig	2.0.1	codetools	0.2-15
devtools	1.13.5	plyr	1.8.4	compiler	3.4.4
dichromat	2.0-0	praise	1.0.0	datasets	3.4.4
digest	0.6.15	R.cache	0.13.0	foreign	0.8-69
emdi	1.1.3	R.methodsS3	1.7.1	graphics	3.4.4
foreign	0.8-69	R.oo	1.21.0	grDevices	3.4.4
ggplot2	2.2.1	R.rsp	0.42.0	grid	3.4.4
git2r	0.21.0	R.utils	2.6.0	KernSmooth	2.23-15
glue	1.2.0	R6	2.2.2	lattice	0.20-35
gridExtra	2.3	RColorBrewer	1.1-2	MASS	7.3-49
gtable	0.2.0	Repp	0.12.16	Matrix	1.2-12
HLMdiag	0.3.1	RcppArmadillo	0.8.400.0.0	methods	3.4.4
hms	0.4.2	RcppEigen	0.3.3.4.0	mgcv	1.8-23
httr	1.3.1	readODS	1.6.4	nlme	3.1-131.1
ineq	0.2-13	readr	1.1.1	nnet	7.3-12
jsonlite	1.5	rematch	1.0.1	parallel	3.4.4
labeling	0.3	reshape2	1.4.3	rpart	4.1-13
laeken	0.4.6	rgeos	0.3-26	spatial	7.3-11
lattice	0.20-35	rlang	0.2.0	splines	3.4.4
lazyeval	0.2.1	RLRsim	3.1-3	stats	3.4.4
lme4	1.1-15	roxygen2	6.0.1	stats4	3.4.4
magrittr	1.5	rprojroot	1.3-2	survival	2.41-3
maptools	0.9-2	rstudioapi	0.7	tcltk	3.4.4
MASS	7.3-49	scales	0.5.0	tools	3.4.4
Matrix	1.2-12	simFrame	0.5.3	utils	3.4.4
memoise	1.1.0	sp	1.2-7		

Table C.1: Packages installed while producing the results presented in this paper.

Part III

Transformations for Achieving Model Assumptions in Linear and Linear Mixed Models

Chapter 5

A guideline of transformations in linear and linear mixed regression models

5.1 Introduction

The linear regression model is perhaps the simplest and most common model used in statistical analysis. The linear mixed regression model is similarly useful for cluster or longitudinal data types. The estimation and inference methods employed with these kinds of models typically rely on a set of assumptions; some of them inherent to the functional form of the model (e.g., linearity), and others related to the nature of the error terms, the response variable, and the covariates (e.g., homoscedasticity). However, empirical data does not always satisfy these assumptions and, therefore, one must decide how to carry on with the analysis. According to Sakia (1992), there are many available options for such cases, which may be summarized as: (i) ignore the violation(s) and proceed; (ii) use a method that allows for the corresponding violation(s); (iii) redesign the model e.g., by properly transforming the data, and (iv) use a distribution-free method. Instead of developing new theories, applying complex methods or extending software functions, using transformations (option (iii)) is a parsimonious way to deal with model assumption violations under both linear and linear mixed regression models. The set of model assumptions that are commonly satisfied by properly transforming the data are normality, homoscedasticity, and linearity. Furthermore, using transformations allows practitioners to apply the most powerful methods available for parametric statistics and to make analysis simpler than otherwise possible. For instance, transformations can allow us to easily get rid of high order terms and work only with first-order linear relationships, which is often preferred in several branches of knowledge (Draper and Hunter, 1969). But how and where are transformations usually used in practice?

The use of transformations has received much attention in the last century in both theoretical knowledge and practical applications (e.g., Edgeworth (1900); Bartlett (1947); Box and Cox (1964)), and is still of great concern in many investigations (e.g., Gurka et al. (2006); Watthanacheewakul (2014)). In the literature of transformations, we find linear, monotonic,

accelerating, and decelerating, power and two-bend transformations, among others. The most discussed type of transformations is the power family, which includes as a particular case both the Box-Cox transformation and the logarithmic function. General overviews about applying transformations under the linear regression model are published by Kruskal (1968); Hoyle (1973); Tukey (1977); Sakia (1992) and Fink (2009). Zarembka (1974a) provides an overview of variable transformations in econometrics. He paid special attention to the problem of heteroscedasticity and illustrated the transformations theory employing elasticity and demand studies. Volatility studies, functional form of demand equations, and economic depreciations have been analyzed mainly using the logarithmic transformation, and also the Box-Cox method (Gemmill et al., 1980; Hulten and Wykoff, 1981; Boylan et al., 1982; Goncalves and Meddahi, 2011). Hossain (2011) gives an analytical review in economic sciences about the importance of the Box-Cox transformation regarding estimation, model selection, and testing. In education, social, biological, and ecological studies, the logarithm is certainly the most relevant transformation and the Box-Cox is also becoming a standard method for variable transformations in these fields (Buchinsky, 1995). In the medical sciences, special attention is paid to dealing with non-normal data (Bland and Altman, 1996). Snedecor and Cochran (1989); Sokal and Rohlf (1995); Keene (1995); Zar (1999) and Armitage et al. (2008) give an introductory literature for medical researches about using transformations, focusing on the logarithmic, Box-Cox, square root, and arcsine transformations. Since biological and medical studies often use longitudinal data, linear mixed regression models for repeated measures analysis are commonly applied (Miller, 2010). In order to deal with model assumption violations under these models, the logarithmic and Box-Cox transformations are preferred (Gurka et al., 2006; Maruo et al., 2017). Furthermore, renowned applications of the Box-Cox transformation in this context are described in Solomon (1985); Piepho and McCulloch (2004); Gurka et al. (2006) and Lo and Andrews (2015).

As we can see, the literature of transformations in theoretical statistics and practical case studies is very rich. However, some important considerations for using them in linear and linear mixed regression models are still broadly discussed: for example, at which stage of the analysis a transformation should be applied, which transformation is suitable for a specific problem and how the results should be interpreted. Practitioners often automatically and routinely apply transformations without considering the questions mentioned above. For this purpose, the present work extends the work of Medina (2017) and proposes a framework that seeks to help the researcher to decide if and how a transformation should be applied in practice. It combines a set of pertinent steps, tables, and flowcharts that guide the practitioner through the analysis of transformations in a friendly and practical manner. This guideline is structured as follows:

- Defining relevant assumptions depending on the research goals
- Choosing a suitable transformation and estimation method according to model assumption violations
- Providing a proper inference analysis and interpreting model results more carefully

Furthermore, the paper points out briefly a selection of special issues that need to be considered when using transformations. To the best of our knowledge, none of the existing reviews for

transformations provides such a comprehensive overview of transformations in the context of linear and linear mixed regression models, as well as developing a practical guideline for researchers.

The remainder of this paper is structured as follows. Section 5.2 guides the reader through the steps of the framework. Each transformation and estimation method is introduced to its corresponding model assumption. Section 5.3 discusses further issues that can arise in modeling and how these interact with the transformations. We conclude the paper in Section 5.4.

5.2 Transformations step framework

*“Although we often hear that data speak for themselves,
their voices can be soft and sly.”*

—Frederick Mosteller

5.2.1 Choose the model and be aware of the corresponding assumptions

Linear regression models are one of the most widely used statistical methods in most branches of knowledge, in particular, the social and natural sciences. It can be expressed in a general form:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, \quad e_i \stackrel{iid}{\sim} N(0, \sigma_e^2), \quad (5.1)$$

where y_i is the target variable defined for the i th individual, with $i = 1, \dots, n$; \mathbf{x}_i^\top is a vector containing deterministic auxiliary information with dimension $1 \times (p + 1)$ and \mathbf{X} would be the corresponding $n \times (p + 1)$ matrix where p is equal to the number of predictors; $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients defined as $\boldsymbol{\beta}^\top = (\beta_0, \dots, \beta_p)$ and e_i is the unit-level error term.

In social, behavioral, educational, and medical sciences, data is commonly hierarchically collected, for instance, as a clustered or longitudinal design (Raudenbush and Bryk, 2002). To appropriately take this type of data structure into account, the so-called linear mixed regression models are typically used. These models, handled as a special extension of the linear regression model, contain additional random-effects depending on the case study and following Laird and Ware (1983), they can be written as

$$\mathbf{y}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{Z}_j \mathbf{u}_j + \mathbf{e}_j, \quad (5.2)$$

where \mathbf{y}_j is a $n_j \times 1$ vector of the dependent variable, n_j is the sample size in each cluster j with $j = 1, \dots, m$ cluster, \mathbf{X}_j is a $n_j \times (p + 1)$ matrix, $\boldsymbol{\beta}$ is the $(p + 1) \times 1$ vector of regression coefficients, \mathbf{Z}_j is the $n_j \times (q + 1)$ matrix with $(q + 1)$ random effects, \mathbf{u}_j is a $(q + 1) \times 1$ vector of random effects and \mathbf{e}_j is the vector of residuals of size $n_j \times 1$. The distribution of the

random effects is given by

$$\mathbf{u}_j \sim N(\mathbf{0}, \mathbf{G}), \quad \text{where } \mathbf{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \dots & \sigma_{0q} \\ \sigma_{10} & \sigma_1^2 & \dots & \sigma_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \dots & \sigma_q^2 \end{bmatrix},$$

and the residuals are distributed with $e_j \sim N(\mathbf{0}, \mathbf{R})$ with $\mathbf{R} = \mathbf{I}_{n_j} \sigma_e^2$ where \mathbf{I}_{n_j} is the $n_j \times n_j$ identity matrix and σ_e^2 is the residual variance. The random effects \mathbf{u}_j and the residuals e_j are assumed to be independent.

Typically the set of assumptions upon which these models rely can be summarized as below:

- (i) The error terms are normally distributed.
- (ii) The error terms have (conditional) homoscedastic variances.
- (iii) The response variable and explanatory variables have a linear and an additive relationship.
- (iv) The error terms are (conditionally) independent.
- (v) The error terms have (conditional) mean equal to zero.

Two potential problems that will also be taken into account are multicollinearity and outliers. However, these are not listed as assumptions for these regression models, since they are not seen as theoretical constraints (Barry, 1993). As we shall discuss in more detail below, if any of these assumptions is violated, estimations, predictions, and scientific insights produced by the linear and linear mixed regression models may be inefficient or, in some cases, severely biased and misleading. This work mainly focuses on the relevance of assumptions (i) - (iii). For readers interested in the assumptions (iv) and (v), discussions, diagnostics and potential solutions are presented in econometric books such as Johnston and DiNardo (1972) and Spanos (1986).

5.2.2 Choose a suitable transformation that addresses assumption violations

The usage of data transformations is directed towards a twofold aim: to create a useful metric or to improve model regression assumptions. For the first aim, linear transformations help in the following ways: information can be easier to understand (e.g., percentage); standardization can be applied in order to change the scale (e.g., covariances into correlation); and a shift can be added to the set of points to make variables positive. Furthermore, these can be useful when transforming qualitative ordinal data into a more convenient and continuous scale, for which normal scores are recommended (for further details see Hoyle, 1973; Fink, 2009). However, such linear transformations do no attempt to correct violations of the regression model assumptions presented in Section 5.2.1. A linear transformation will change only the intercept of the regression equation. For instance, using this type of transformation does not help to linearize

non-linear relations (Brown, 2015). In this work, we focus on transformations that attempt to correct violations of the assumptions of the linear and linear mixed regression model. These non-linear transformations are monotonic and shrink or stretch a topological space in an inhomogeneous way. That is, the order of the points lying on this space remain unchanged, but the relative distance between them will be altered (Cohen et al., 2014).

For defining such a transformation, the following notation is used consistently through the present work. We denote y as the response variable with expected value denoted by $E(y) = \mu_y$ and variance by $V(y) = \sigma_y^2$. For a single untransformed observation we use y_i, y_{ij} where an additional symbol * denotes that the observation is transformed. The untransformed vector of the response variable is defined as \mathbf{y} . Furthermore, $\mathbf{y}^*(\Theta)$ represents the vector of the transformed observations of the response variable and Θ represents the set of parameters upon which the transformation depends. The transformation parameter is generally denoted by λ , but it depends on the functional form of the transformation. Some transformations also include additional parameters. The relationship between original and transformed data is denoted by $T(y) = y^*$.

For this section, the following structure is used: we describe the model assumption and its relevance, we introduce assessment tools to check its fulfillment, we mention alternative methods to transformations, and we discuss the range of possibilities using transformations with corresponding estimation methods.

5.2.2.1 Transformations to achieve normality

Why is the normality assumption important?

The fulfillment of the normality assumption is usually twofold: it builds confidence intervals and for computing statistical tests and appropriately uses the percentage points of customary tables of χ^2, t, F distributions. When this assumption is not fulfilled, practical problems can arise; as for estimation, the ordinary least square method does not provide best estimators in terms of efficiency, in case the true distribution of the error term is skewed or has heavy tails. When the interest lies in inference hypothesis testing, such as a t-test for significance of the coefficients, the results of this test seem to be fairly robust for large enough samples. However, its power may be somewhat affected when, for instance, the true distribution has heavy tails, as σ_e^2 is very sensitive to values at the tails of the distribution (Wilcox, 2005). The most common departures from normality are skewed, heavy-tailed, and light-tailed distributions. Additionally, human errors can contribute to the presence of non-random aspects which lessen the strength of the assumption that the error term is normally distributed (Zeckhauser and Thompson, 1970). Some papers related to the consequences when Gaussian assumptions are not satisfied are published by Fisher (1922b); Pearson (1931); Bartlett (1935); Hey (1938); Finney (1941), and Cochran (1947).

How can we check the normality assumption?

Due to the importance of the normality assumption, many methods have been developed to check its validity: visual methods such as the normal probability plot of the residuals (Cham-

bers et al., 1983), histograms, and probability plots. The normal probability plot, also known as normal scores plot, quantile-quantile (Q-Q) plot, quantile comparison plot or rankit plot can be useful for comparing two probability distributions in terms of the location, scale, and skewness parameters (Weisberg, 1980; Bock, 1985; Fox, 1997; Hutcheson and Sofroniou, 1999; Johnson, 2009). The histogram is a standard visualization of the empirical distribution form. The probability plot, also known as probability-probability (P-P) plot or percent-percent plot, is suitable for analyzing the skewness of a distribution, by plotting two cumulative distribution functions. Numerical analysis of the distribution moments, such as skewness and kurtosis, is a common rule-of-thumb for checking the normality assumption. The skewness and kurtosis for a normal distribution are equal to zero and three, respectively. Therefore, a comparison with this distribution is often made in practice. Additionally, normality tests such as the Kolmogorov-Smirnov test (Smirnov, 1948), Anderson-Darling test (Anderson and Darling, 1954) and the Shapiro-Wilk test (Shapiro and Wilk, 1965) are also widely used.

What are the alternative methods to overcome non-normality?

If any of the aforementioned techniques suggests that the data is not normally distributed, we could move to non-normal methods or redesign the model. In this case, there are some typically recommended solutions. The first method and perhaps the most common one is to allow a more flexible model where the conditions imposed over the error term and independent variables can be relaxed. This method is known as the generalized linear or generalized linear mixed model (Nelder and Wedderburn, 1972). A second solution is to work with more robust tests such as the Kruskal-Wallis (Kruskal and Wallis, 1952) or the Levene's test (Levene, 1960). Robust and more efficient estimators have been studied when the error term is not normally distributed (see, for instance, Huber, 1964). Among these approaches, we find non-parametric maximum likelihood theory (Aitkin, 1999; Agresti et al., 2004; Litière et al., 2008), more flexible parametric distributions (Peng Zhang and Greene, 2008), marginalized mixed effects models (Heagerty and Zeger, 2000), and h-likelihood approaches that can be adapted to fit different distributions (Lee et al., 2004). Also possible are methods based on mixtures of normal distributions (Lesaffre and Molenberghs, 1991) and "smooth" non-parametric fits (Zhang and Davidian, 2001).

How can transformations help to improve normality?

The use of transformations is considered as a parsimonious alternative to complex methodologies when dealing with the departure from normality, a feature seldom observed in raw data. A significant part of the effort put into transformations has been focused on achieving approximate normally distributed errors. To ensure normality, it is common to use a proper one-to-one transformation on the target variable (Thoni, 1969; Hoyle, 1973). A standard practice in applied work is transforming the target variable by computing its logarithm. That means using a transformation of the form $\log(y)$. Due to its effectiveness in turning highly right-skewed or log-normal distributions into more symmetrical ones, it is commonly used in practice for this purpose. Furthermore, the logarithmic transformation is used in parallel for achieving normality, homoscedasticity, and linearity (Bartlett and Kendall, 1946; Bartlett, 1947; Anscombe, 1948; Kleczkowski, 1949; Moore, 1958). However, the ease of its use and its popularity often

induce an imprudent application (Feng et al., 2014). One drawback of the logarithmic transformation is the lack of ability to deal with negative values. Thus, some adjustments based on the logarithm have been proposed. A simple shifted version includes a fixed term s such that $y + s > 0$. The logarithmic transformation is often recommended when dealing with substantially positive skewness. For a left-skewed distribution, the log neg transformation is suggested. It includes a fixed parameter p for which every observation of the target variable is subtracted so that the smallest score is 1 (Tabachnick and Fidell, 2007). Furthermore, the generalized logarithm, also known as the glog transformation allows for negative values, but it is recommended for low values rather than high ones (Durbin et al., 2002; Huber et al., 2003). Even though it is suitable for correcting non-normality, it is more widely used as a variance stabilizing transformation. Another transformation used particularly for dealing with non-negative variables such as the non-central chi-square is suggested by Moschopoulos (1983). He bases his work on the theory developed by Jensen and Solomon (1972), including the moments of the distribution as transformation parameters. Square roots and inverse transformations are commonly used for dealing with right-skewed distributions (Bartlett, 1937). The square root is also used for dealing with data having zero inflation problems or containing extremely small values. The cube-root transformation, also known as the Wilson-Hilferty (Wilson and Hilferty, 1931) transformation, is particularly suitable for symmetrizing gamma-distributed data forms. The exponential, square, and cube root transformations are commonly used for negative skewed data. A quasi generalization of this problem is made in practice in the transformation exponent: right-skewed distributions tend to be more symmetrical by applying a transformation with an exponent smaller than one, and left-skewed distributions, with an exponent greater than one (Hoaglin et al., 2000). When comparing the square-root transformation with the logarithm, Garson (2012) states that the latter is more useful in case symmetry in the central distribution is needed. Meanwhile the square root is suggested in case symmetry in the tails is more important. Finally, in the case of negative skewness, the reciprocal transformations may be useful as an appropriate variance stabilizing transformation (Hoyle, 1973) for certain distributions.

The transformations mentioned so far have in common that they do not adjust to the underlying data. To find a data-driven transformation, an adjustment is done by including a data-driven transformation parameter, denoted by λ . This parameter should be estimated and this estimate changes according to the data, the assumption violations or to a specific researcher criteria. For instance, an advanced log-shift opt transformation used in practice (e.g., Feng et al., 2016) includes an optimal transformation parameter as follows $y^*(\lambda) = \log(y + \lambda)$. Tukey (1957) proposed a family of power transformations based on monotonic functions. The general form of this family is defined as: y^λ if $\lambda \neq 0$ and $\log(y)$ if $\lambda = 0$. The power transformations are also commonly denoted as single- or one-bend transformations (Box and Cox, 1964; Montgomery, 2008; Fink, 2009; Cohen et al., 2014). To avoid the discontinuity at $\lambda = 0$, Box and Cox (1964) modified this family. The straightforward manner in which the interpretation of this parameter is made makes the Box-Cox method one of the most widely used transformations. For instance, when $\lambda = -1$, it means the reciprocal transformation is needed, $\lambda = 0$ means the logarithmic transformation is recommended, $\lambda = 1/2$ implies the use of the square root and $\lambda = 1$ suggests that no transformation is necessary. The Box-Cox transformation

is the simplest single-bend transformation (Fink, 2009) and is more appropriate when dealing with skewed distributions than symmetric but non-normal distributions. It has been extensively implemented in different branches of knowledge. For detailed information about renowned applications, see Draper and Cox (1969), Mills (1978), Poirier (1978), Machado and Mata (2000), Chen (2002), Chen and Deo (2004) and Yang and Tsui (2004).

Since the Box-Cox transformation is not defined for negative values, the data must be shifted to the positive side by incorporating a shift parameter. This method is known as the shifted power transformation. It overcomes the difficulties encountered in the Box-Cox transformation due to the restriction $y > 0$. This is done by incorporating a constant, denoted by s , for accommodating negative values of the target variable. The parameter s is chosen such that $y + s > 0$. Moore (1957) studies the benefits of adding this shift parameter in the power family of transformations. However, Hill (1963); Atkinson (1987) and Yeo and Johnson (2000) state that shifting the data is not always an optimal way to deal with negative values. Different modifications have been proposed in the literature to address this issue. The first proposal to avoid this difficulty was made by Manly (1976), who proposes the Manly transformation, an exponential power transformation family. This transformation family is considered to nearly normalize unimodal skewed distributions, but it is not suitable for bimodal or U-shape distributions. In case the data also presents a symmetric but non-normal error distribution, the modulus power transformation proposed by John and Draper (1980) should be used. It can manage negative values and is claimed to be effective for somewhat symmetrical or bimodal distributions. In the same way, the neglog transformation, proposed by Whittaker et al. (2005) is developed especially to deal with negative values. In order to avoid the non-negativity restriction of the Box-Cox transformation, Bickel and Doksum (1981) introduced the Bickel-Doksum power transformation which is defined on the whole real line. This transformation is especially useful for handling kurtosis rather than skewness, in particular for leptokurtic and platykurtic distributions. However, as Yeo and Johnson (2000) point out, one should avoid the use of this transformation when dealing with skewed data that takes negative and positive values. As another alternative to the Box-Cox transformation, Kelmansky et al. (2013); Kelmansky and Ricci (2017) recently proposed an extension of the glog transformation, also known as gpower transformation. It allows for negative values, heavier tails and peaked sample modes (Tsai et al., 2017). The work of MacKinnon and Magee (1990) proposes a scale-invariant family of transformations, which deals with variables with zero or negative values.

Zwet (1964) emphasizes that for reaching near symmetry when the response variable has positive and negative values, the transformation should be concave. One could say that a transformation has the quality of reducing left-skewness if such a transformation is non-decreasing convex or upward bending, and a transformation is needed to symmetrize right-skewness if such a transformation is non-decreasing concave or downward bending. Under this motto, different transformations have been proposed for kurtosis adjustments in order to deal with non-normality. This is also achieved by the convex-to-concave Yeo-Johnson transformation (Yeo and Johnson, 2000) for different ranges of λ . The transformation is convex in y for $\lambda > 1$, and concave for $\lambda < 1$. Nevertheless, this transformation is not suitable when data has a platykurtic, leptokurtic or bimodal form. Analogously, the power transformations family

is convex in case $\lambda > 1$ and concave when $\lambda < 1$. Following Tsai et al. (2017), transformations that are suitable for data with a peaked mode are the signed power (Bickel and Doksum, 1981), the modulus (John and Draper, 1980), the sinh-arcsinh (Jones and Pewsey, 2009), the gpower (Kelmansky et al., 2013) and the inverse hyperbolic sine (Johnson, 1949; Burbidge et al., 1988). The signed transformation is convex-concave as the outcome variable changes the sign, which is an effect that is difficult to predict. Therefore, it is recommended to use it for a kind of symmetric distribution in order to deal with the kurtosis, rather than skewness (Zwet, 1964; Oja, 1981).

Another difficulty of the Box-Cox transformation is the truncation on the transformation parameter determined by λ . If λ is positive, y^* has an upper-bound at $\frac{-1}{\lambda}$ and if λ negative y^* has also a lower-bound at $\frac{-1}{\lambda}$. Thus, achieving exact normality is not possible if $\lambda \neq 0$. In order to deal with this problem, Yang (2006) recently proposed the dual power transformation. It is defined only for strictly positive values. In the case that the outcome variable is bounded above as well as below, the previous transformations are not suitable. Therefore, the appropriate transformation based on an interval $[0, b]$ is the folded-power transformation (Mosteller and Tukey, 1977; Atkinson, 1982). However, if the outcome scores are close to 0 or b the behavior would be like the Box-Cox transformation (Cook and Weisberg, 1982). A practical application of the shifted version of the dual transformation is shown, for instance, in Rojas-Perilla et al. (2017).

Besides the power transformations presented above, the multi-parameter transformation families have been suggested in order to estimate different transformation parameters, accounting for scale, location, and shape (skewness and tailweight). For this purpose, Johnson (1949) proposes three normalizing transformations, which include shape, scale, and location parameters, where a system of curves represents the empirical distributions (Edgeworth, 1900). Furthermore, for continuous empirical forms, this method has the particular advantage that many distributions can be fitted into the system, which delivers a high flexibility that can be advantageous for dealing with complicated data sets (George, 2007). As a special case of the Johnson transformation, the one-parametric inverse hyperbolic sine is suitable for dealing with negative and positive values (Burbidge et al., 1988). This transformation contains the Pearson system of frequency curves (Pearson, 1894). These curves properly represent data which exhibit departures from normality or with considerable skewness, that means non-normal forms. In contrast, the sinh-arcsinh transformations are applied for heavy-tailed and light-tailed distributions (Jones and Pewsey, 2009).

As mentioned before, in many branches of knowledge, cross-sectional data are widely used. However, little attention has been paid to the study of techniques in the literature of linear mixed regression models, which assess or improve the validity of the multiple distributional assumptions by departures from normality of the error terms expressed in Equation 5.2. In order to improve the assumptions of the model by parametrically transforming the outcome variable in linear mixed regression models, single-bend transformations, such as the logarithmic and square root transformations, have been applied in particular case studies (McCulloch and Neuhaus, 2001; Piepho and McCulloch, 2004; West et al., 2007; Lo and Andrews, 2015). Solomon (1985) and Lipsitz et al. (2000) have furthermore studied the application of the Box-

Table 5.1: Transformations for achieving normality.

Transformation	Source	Formula	Support	N	H	L
Log	Tukey (1977)	$\log(y)$	$y > 0$	✗	✗	✗
Log (shift)	Box and Cox (1964)	$\log(y + s)$	$y \in \mathbb{R}$	✗	✗	✗
Log neg	Tabachnick and Fidell (2007)	$\log(p - y)$	$y \in \mathbb{R}$	✗	✗	✗
Glog	Durbin et al. (2002)	$\log(y + \sqrt{y^2 + 1})$	$y \in \mathbb{R}$	✗	✗	✗
Moschopoulos	Moschopoulos (1983)	$\left(\frac{y+a}{\mu}\right)^b$	$y > 0$	✗		
Square Root	Bartlett (1937)	\sqrt{y}	$y > 0$	✗	✗	
Square root neg	Tabachnick and Fidell (2007)	$\sqrt{p - y}$	$y \in \mathbb{R}$	✗	✗	
Wilson-Hilferty	Wilson and Hilferty (1931)	$y^{1/3}$	$y \in \mathbb{R}$	✗	✗	
Reciprocal	Tukey (1977)	$\frac{1}{y}$	$y \neq 0$	✗	✗	
Log-shift opt	Feng et al. (2016)	$\log(y + \lambda)$	$y \in \mathbb{R}$	✗	✗	✗
Folded	Mosteller and Tukey (1977)	$y^\lambda - (1 - y)^\lambda$ if $\lambda \neq 0$.	$y > 0$	✗	✗	
Box-Cox	Box and Cox (1964)	$\begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗	✗	✗
Box-Cox (shift)	Box and Cox (1964)	$\begin{cases} \frac{(y+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y + s) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	✗
Manly	Manly (1976)	$\begin{cases} \frac{e^{\lambda y} - 1}{\lambda} & \text{if } \lambda \neq 0; \\ y & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Modulus	John and Draper (1980)	$\begin{cases} \text{Sign}(y) \frac{(y +1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \text{Sign}(y) \log(y + 1) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		
Neglog	Whittaker et al. (2005)	$\text{Sign}(y) \log(y + 1)$	$y \in \mathbb{R}$	✗	✗	
Bickel-Docksum	Bickel and Doksum (1981)	$\frac{ y ^\lambda \text{Sign}(y) - 1}{\lambda}$ for $\lambda > 0$	$y \in \mathbb{R}$	✗	✗	
Gpower	Kelmansky et al. (2013)	$\begin{cases} \frac{(y + \sqrt{y^2 + 1})^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y + \sqrt{y^2 + 1}) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		
Mackinnon-Magee	MacKinnon and Magee (1990)	$\frac{h(\lambda y)}{\lambda}$	$y \in \mathbb{R}$	✗		✗
Yeo-Johnson	Yeo and Johnson (2000)	$\begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0; \\ \log(y + 1) & \text{if } \lambda = 0, y \geq 0; \\ \frac{(1-y)^{2-\lambda} - 1}{\lambda - 2} & \text{if } \lambda \neq 2, y < 0; \\ -\log(1 - y) & \text{if } \lambda = 2, y < 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Dual	Yang (2006)	$\begin{cases} \frac{(y^\lambda - y^{-\lambda})}{2\lambda} & \text{if } \lambda > 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗		
Tukey	Tukey (1957)	$\begin{cases} y^\lambda & \text{if } \lambda \neq 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗	✗	
Johnson	Johnson (1949)	$\kappa + \nu h\left(\frac{y - \xi}{\eta}\right)$	$y \in \mathbb{R}$	✗	✗	
Sinh-arcsinh	Burbidge et al. (1988)	$\sinh[\theta \sinh^{-1}(y - \gamma_1)]$	$y \in \mathbb{R}$	✗		

Note: Normality, homoscedasticity, and linearity are denoted as N,H,L, respectively. Additional to the notation that is used throughout the paper, for some transformations further parameters need to be defined. The parameters s and p are fixed parameter and chosen such that the smallest score is equal to 1. In the Moschopoulos transformation μ is the first moment of the distribution, and a and b are determined from the first three moments of the distribution. The known fixed values that work for this transformation are $b = 1/3$ (Wilson and Hilferty, 1931) and $b = 1/2$ (Fisher, 1922a). In the transformation by MacKinnon and Magee (1990), $h(\cdot)$ is a monotonically increasing function that satisfies the following properties: $h(0) = 0$, $h'(0) = 1$ and $h''(0) \neq 0$. One common function is defined as $h(\cdot) = \sinh^{-1}(y)$. According to Johnson (1949), η and ν are the scale parameters and κ and ξ the location parameters. $h(\cdot)$ is a monotonic function of y . In the sin-arcsinh transformation, $\gamma_1 \in \mathbb{R}$ represents the skewness parameter and $\theta > 0$ controls the tail weight.

Cox transformation to cover all linear mixed regression models and some longitudinal datasets, while the work of Gurka et al. (2006) formally extended the use of the Box-Cox method for these kinds of models.

Finally, as Box and Cox (1964) state, several transformations are suitable to improve not only one model assumption, but many. This is also expressed in Table 5.1, which contains transformations that help to achieve normality. Additionally, it is indicated which further assumption can be often improved by these transformations. We exhaustively examine the literature on transformations and present it in Table 5.1 and subsequent tables as a condensed version of the research work.

How can we estimate the transformation parameter to normality?

In addition to the selection of a suitable transformation, different methodologies for the estimation parameter have been introduced. The estimation method partly depends on which model assumption we want to enforce. Please notice that some of the estimation methods are, so far, only developed for the Box-Cox transformation. In general, the approaches for estimating the optimal transformation parameter to normality are classified in maximum likelihood-based approaches (A), analytical considerations (B), robust adaptations (C), and Bayesian approaches (D). The methods are described below and the mathematical formulation is presented in detail for these ones, which are more commonly applied.

A: Maximum likelihood-based approaches

A.1: Maximum likelihood (ML) approach

The ML-based method is also known as the profile log-likelihood approach. It is the most commonly cited approach under the linear regression model and is described in detail in Box and Cox (1964). It has been studied by Draper and Cox (1969); Andrews (1971); Atkinson (1973); Carroll (1980) and Bickel and Doksum (1981). The goal is to find the transformation parameter λ for which the expected value $E[\mathbf{y}^*(\lambda)]$ is equal to $\mathbf{X}\boldsymbol{\beta}$ meeting the model assumptions listed in the previous chapter. If the normality assumption $y_i^*(\lambda) \sim N(\mathbf{x}_i\boldsymbol{\beta}, \sigma_e^2)$ is fulfilled, the probability density function for $y_i^*(\lambda)$ is written as

$$f(y_i^*(\lambda)) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp \left\{ -\frac{(y_i^*(\lambda) - \mathbf{x}_i\boldsymbol{\beta})^2}{2\sigma_e^2} \right\}. \quad (5.3)$$

The probability density function for the untransformed observations, and thus the likelihood for the whole (transformed) model in relation to those observations, is computed as the likelihood of Equation 5.3 multiplied by the Jacobian of the transformation, explicitly:

$$L(\mathbf{y}, \lambda \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}} \exp \left\{ -\frac{(\mathbf{y}^*(\lambda) - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y}^*(\lambda) - \mathbf{X}\boldsymbol{\beta})}{2\sigma_e^2} \right\} J(\lambda, \mathbf{y}),$$

where

$$J(\lambda, \mathbf{y}) = \prod_{i=1}^n \left| \frac{\partial y_i^*(\lambda)}{\partial y_i} \right|$$

is the Jacobian of the transformation from y to $y^*(\lambda)$ and θ are the unknown parameters β and σ_e^2 . This property comes from the transformation theorem defined as:

Theorem 1 (Transformation theorem). *Let y be a continuous random variable with density function $f(y)$, taking values in \mathbb{R}^n . Let $T(y) = y^*$ a continuous transformation $T(y) : \mathbb{R}^n \rightarrow \mathbb{R}^n$, for which the inverse $T^{-1}(y^*)$ is also continuous. Suppose that the inverse of the transformation is differentiable for all values of \mathbb{R}^n and the Jacobian is not equal to zero. Then $f_{T(y)}(y)$, the density function of the transformed target variable, is given by:*

$$f_{T(y)}(y) = f\left[T^{-1}(y^*)\right]|J(y)|.$$

The maximum likelihood estimates are found in two stages. First, for fixed λ , the estimates for β and σ_e^2 are computed. When the Jacobian does not depend on β or σ_e^2 this is the likelihood for a least-square problem with response $y^*(\lambda)$. Hence

$$\hat{\beta}(\lambda) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}^*(\lambda),$$

and

$$\hat{\sigma}_e^2(\lambda) = \frac{\mathbf{y}^*(\lambda)^\top \mathbf{A} \mathbf{y}^*(\lambda)}{n} = \frac{S(\lambda)}{n},$$

where $\mathbf{A} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ and $S(\lambda)$ is the residual sum square in the transformed model. Holding λ as fixed and substituting $\hat{\beta}(\lambda)$ and $\hat{\sigma}_e^2(\lambda)$ into the logarithm, we obtain, apart from a constant,

$$l_{max}(\lambda) = -\frac{n}{2} \log \hat{\sigma}_e^2(\lambda) + \log J(\lambda, \mathbf{y}). \quad (5.4)$$

The λ that maximizes the profile log-likelihood in Equation 5.4 will be selected. For the underlying optimization process by using the ML estimation method, the Newton-Raphson iterative procedure and its modifications are commonly used (Nelder and Mead, 1965; Lagarias et al., 1998).

A.2: Restricted maximum likelihood estimation method (REML)

As mentioned before, little attention has been paid in the literature to the study of data-driven transformations for linear mixed regression models: in particular, the improvement of the validity of model assumptions by departures from normality of both sources of randomness and the transformation parameter estimation methods are still under research. The work of Gurka et al. (2006) extends the use of the Box-Cox transformation under maximum likelihood theory for the estimation of the transformation parameter to the linear mixed regression models theory.

For the estimation of λ under the linear mixed regression model presented in Equation 5.2 and described in Gurka et al. (2006), we assume that the vectors \mathbf{y}_j^* are independent and normal distributed for some unknown λ as follows:

$$\mathbf{y}_j^*(\lambda) \sim N(\boldsymbol{\mu}_j, \mathbf{V}_j) \quad \text{for } j = 1, \dots, m,$$

with

$$\boldsymbol{\mu}_j = \mathbf{X}_j \beta \quad \text{and} \quad \mathbf{V}_j = \mathbf{Z}_j \mathbf{G} \mathbf{Z}_j^\top + \mathbf{R}.$$

Let $J(\lambda, \mathbf{y})$ be the Jacobian of the transformation from y to $y_j^*(\lambda)$, defined as

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{j=1}^m \prod_{i=1}^{n_j} \left| \frac{dy_{ij}^*(\lambda)}{dy_{ij}} \right| \\ &= \prod_{j=1}^m \prod_{i=1}^{n_j} y_{ij}^{\lambda-1}, \end{aligned}$$

then, the log-likelihood function in relation to the original observations is obtained by multiplying the normal density by $J(\lambda, \mathbf{y})$ as:

$$\begin{aligned} l_{\text{ML}}(\mathbf{y}, \lambda | \boldsymbol{\theta}) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^m \log |\mathbf{V}_j| \\ &\quad - \frac{1}{2} \sum_{j=1}^m [\mathbf{y}_j^*(\lambda) - \mathbf{X}_j \hat{\boldsymbol{\beta}}]^\top \mathbf{V}_j^{-1} [\mathbf{y}_j^*(\lambda) - \mathbf{X}_j \hat{\boldsymbol{\beta}}] + \log J(\lambda, \mathbf{y}). \end{aligned}$$

The maximization process of $l_{\text{ML}}(\boldsymbol{\theta})$ leads to ML estimators of the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_e^2, \mathbf{G})$. However, the REML theory is recommended when more accurate estimators of the variance components are needed (Verbeke and Molenberghs, 2000). This function is calculated by maximizing the ML of a set of error contrasts stemming from the fixed effects design matrix (Gurka et al., 2006). As a result, the REML function, in which the maximum possible number of linearly independent contrasts is $n - p$ (Harville, 1974), does not depend on $\boldsymbol{\beta}$ as follows:

$$\begin{aligned} l_{\text{REML}}(\mathbf{y}, \lambda | \boldsymbol{\theta}) &= -\frac{n-p}{2} \log(2\pi) + \frac{1}{2} \log \left| \sum_{j=1}^m \mathbf{X}_j^\top \mathbf{X}_j \right| - \frac{1}{2} \sum_{j=1}^m \log |\mathbf{V}_j| \\ &\quad - \frac{1}{2} \log \left| \sum_{j=1}^m \mathbf{X}_j^\top \mathbf{V}_j^{-1} \mathbf{X}_j \right| - \frac{1}{2} \sum_{j=1}^m [\mathbf{y}_j^*(\lambda) - \mathbf{X}_j \hat{\boldsymbol{\beta}}]^\top \mathbf{V}_j^{-1} [\mathbf{y}_j^*(\lambda) - \mathbf{X}_j \hat{\boldsymbol{\beta}}] \\ &\quad + n(\lambda - 1) \log(\bar{y}), \end{aligned}$$

in which \bar{y} , is the geometric mean, defined as

$$\bar{y} = \left(\prod_{j=1}^m \prod_{i=1}^{n_j} y_{ij} \right)^{1/n}.$$

Bickel and Doksum (1981) studied the estimation properties of the parameters while using the Box-Cox transformation, whereby the inference about $\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2$ is conditioned on $\lambda = \hat{\lambda}$. They conclude that the asymptotic marginal unconditional variance of $\hat{\boldsymbol{\beta}}$ can be inflated for a fixed λ . The standard solution to this problem is to include the geometric mean of the response variable in the denominator of the Box-Cox transformation $\frac{y_{ij}^*(\lambda)}{J(\lambda, \mathbf{y})^{1/n}}$, which converts it in a scaled transformation $Z(\lambda)$, whereby the unit is preserved and the interpretation is simplified, due to the fact that the units do not change as λ changes and the conditional variance of $\boldsymbol{\beta}$ is reduced. The Jacobian of this transformation is equal to one and the ML theory can be used for

the linear mixed regression model. It is defined as follows:

$$Z(\lambda) = \begin{cases} \frac{y_{ij}^\lambda - 1}{\bar{y}^{\lambda-1}} & \text{if } \lambda \neq 0; \\ \bar{y} \log(y_{ij}) & \text{if } \lambda = 0, \end{cases}$$

for $y_{ij} > 0$. Gurka et al. (2006) recommend this scaled transformation in order to take advantage of procedures for estimating λ already computationally implemented.

B: Analytical considerations

Other analytical considerations have also been proposed in the literature as alternatives to ML-based methods. It consists of the use of distances or divergence measures, fit tests, and distribution moments (Hernandez and Johnson, 1980; Yeo and Johnson, 2000; Vélez et al., 2015). These approaches have also been studied in the context of linear mixed regression models (see e.g., Rojas-Perilla et al., 2017). For some multiparameter transformations, such as the Johnson's System (Johnson, 1949), the method of moments of percentile points is proposed (George, 2007; Forbes et al., 2011). It is based on a simple selection rule introduced by Slifker and Shapiro (1980). Therefore, Chung et al. (2007) recommend the method of percentiles over the profile log-likelihood due to its simplicity. The hyperbolic power transformation is another example of a multi-parameter transformation. For this, the matching quantile approach (Tsai et al., 2017) is used for estimating the transformation parameters. Finally, in case the outcome variable is truncated, Poirier (1978) introduced a methodology as alternative of the ML method.

B.1 Estimators based on goodness of fit tests

In simple words, a goodness of fit test compares the empirical distribution, g , of a random sample against a theoretical distribution, f . Typically, a null hypothesis, H_0 , is tested that assumes that f and g are statistically equal. If the hypothesis is rejected, we say that there is ground for believing that the sample is not f distributed. If we fail to reject H_0 , the hypothesis that the sample is f distributed cannot be discarded. In the frame of the present work, f is the density function of the normal distribution. The goodness of fit tests can be employed to estimate the transformation parameter. The main idea is to maximize the statistic of such tests. Rahman (1999) employs the Shapiro-Wilk test. Rahman and Pearson (2008) make use of the Anderson-Darling test. Both focus on the Box-Cox transformation and use the Newton-Raphson algorithm to estimate the transformation parameter. However, these methods can also be applied to all one-parameter transformations mentioned in the present work. Yang and Abeyasinghe (2003) make use of two score tests to determine transformation parameter for the Box-Cox transformation. Applications for multiple parameters transformations such as the Johnson transformation need to be further studied. Asar et al. (2017) extend the work of Rahman (1999) and Rahman and Pearson (2008) by utilizing seven goodness of fit tests, proposing a new algorithm. For a more detailed description about their method see Asar et al. (2017). Ruppert and Aldershof (1989) introduce an estimator for λ , σ_e^2 and β based on a test which depends on the correlation of the fitted values with the squared residuals. Other versions of this type of estimator are based on the Levene's test and Anscombe test. Finally,

the work of Vélez et al. (2015) makes a selection of different normality type tests, classified in regression/correlation-, empirical distribution function-, and measure of moments-based tests. They develop a grid-search method for choosing the transformation where the combined p -value is the highest.

B.2: Estimators based on distribution moments: skewness and kurtosis

Skewness and kurtosis are major characteristics of the shape of distributions (Rosenthal, 2011). The former is a measure of the degree to which a distribution departs from symmetry; if it is negative, the left tail is long and the right short and thick. Positive values of skewness mean the contrary: a large right tail and a stubby left tail. The normal distribution has a skewness equal to zero. For a random variable z with mean μ_z and variance σ_z^2 the skewness is defined as

$$\gamma_1(z|\mu_z, \sigma_z^2) = E\left[\left(\frac{z - \mu_z}{\sigma_z}\right)^3\right].$$

Kurtosis is a measure of the degree of “tailedness” or “peakedness” concerning the normal distribution. A leptokurtic distribution has high kurtosis, which means that the probability of falling in the center is greater compared to that of the normal distribution. In contrast, a platykurtic distribution has more area, and therefore, more probability in the tails. The kurtosis for a standard normal distribution is equal to 3. Typically, the interest lies in the excess of kurtosis, which for the random variable z is defined as follows:

$$\gamma_2(z|\mu_z, \sigma_z^2) = \left[\frac{(E[z - \mu_z])^4}{(E[(z - \mu_z)^2])^2} \right] - 3.$$

Even though the skewness is considered more important than the kurtosis when dealing with model assumption violations (Royston et al., 2011), the optimization can be made for both measures. The parameter of the transformation is then chosen so that the value of skewness or kurtosis for the error term e_i is as close as possible to that of the normal distribution (Carroll and Ruppert, 1987).

$$\hat{\lambda}_{\text{skew}} = \underset{\lambda}{\operatorname{argmin}} |\gamma_1(e_i)|,$$

and

$$\hat{\lambda}_{\text{kurt}} = \underset{\lambda}{\operatorname{argmin}} |\gamma_2(e_i)|.$$

where $\gamma_1(e_i)$ is the skewness and $\gamma_2(e_i)$ denotes the kurtosis of the unit-level error terms.

The parameter could be also selected with the help of a statistical test that accounts for kurtosis or skewness (see, for instance, Gaudard and Karson (2007)). In the context of linear mixed regression models, an additional problem arises as there are two independent error terms to be considered. Therefore, a pooled skewness approach is suggested by Rojas-Perilla et al. (2017), if skewness minimization is chosen as the target criteria. This ensures that the larger

the error term variance is, the more importance its skewness in the optimization has.

B.3 Estimators based on divergence or distance optimization

Only considering skewness may ignore many other properties of the distribution. Hence, a measure describing the distance between two distribution functions as a total might be preferable. A few of these alternatives are based on the minimization of the Kullback-Leibler (KL) divergence, based on Kullback (1997) and described in Yeo and Johnson (2000) and Hernandez and Johnson (1980), and on measures of symmetry as the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) distances (Carroll, 1980; Bickel and Doksum, 1981; Carroll, 1982a; Taylor, 1985). For this method the real distribution of the data needs to be known. The exact formulations of the target measures are given as follows:

$$\hat{\lambda}_{\text{KL}} = \operatorname{argmin}_{\lambda} \int_{-\infty}^{+\infty} f(y^*(\lambda)) \log \left[\frac{f(y^*(\lambda))}{\phi_{\mu, \sigma^2}} \right],$$

with f the probability density function of the transformed target variable $y^*(\lambda)$. ϕ_{μ, σ^2} denotes the probability density function of a normal distribution with mean μ and variance σ^2 .

$$\hat{\lambda}_{\text{KS}} = \operatorname{argmin}_{\lambda} \sup |(F(e^{std}) - \Phi)|,$$

$$\hat{\lambda}_{\text{CvM}} = \operatorname{argmin}_{\lambda} \int_0^1 [F(e^{std}) - \Phi]^2.$$

$F(\cdot)$ denotes the empirical cumulative distribution function (ecdf) estimated on the normalized residuals e^{std} and Φ is the distribution function of a standard normal distribution.

C: Robust adaptations

Draper and Cox (1969) stated that the ML method is robust to non-normal error terms as long as they are reasonably symmetric. It depends on parametric distributional assumptions and it is not robust to outliers. Therefore, different robust adaptations are proposed in the literature. Hinkley (1975, 1977); Hinkley and Runger (1984) and Taylor (1985) introduce and discuss a non-parametric and symmetry-based adaptation method of the ML procedure. This quick-choice method uses a symmetric distribution of the error terms about zero rather than the normal, and is based on an asymmetry measure based on order statistics (Taylor, 1985). It is also known as the Hinkley's quick method or quantile-based method because it studies how the quantiles of the distribution are symmetrically placed about the median. While this approach is not sensitive to outliers and robust in case the interquartile range is used, it is an inefficient method. Another similar quantile-based method for assessing the need of transforming data is suggested in Velilla (1993). Leinhardt and Wasserman (1979) and Emerson and Stoto (1982) propose the symmetrization of the quartiles around the median. However, Cameron (1984) pointed out that the method of Emerson and Stoto (1982) is not suitable for highly skewed data.

In order to access the accuracy of the ML estimator, Carroll (1980) and Bickel and Doksum (1981) propose another robust modification, also studied in Carroll (1982a) and Hinkley and

Runger (1984). It generates a famous controversy in the study of transformations (see Doksum, 1984; Rubin, 1984; Johnson, 1984; Carroll and Ruppert, 1984). They propose a robustification against heavy-tailed distributions in case the normality assumption is not present in the data and the Box-Cox transformation is required. This method is based on the robust estimator defined by Huber (1981), but it is not consistent in terms of mean squared error. Please note that these robust adaptations are made for handling outliers only in the outcome variable and not in the explanatory ones.

In order to find a consistent and efficient non-parametric method, Han (1987) suggests an estimator based on the Kendall's rank correlation (Kendall, 1938). With the aim of covering some heavy tailed distributions, Carroll and Ruppert (1985, 1987, 1988) proposed a robust bounded influence method based on Kruskal and Wallis (1952) to a moderate number of outlying points in the data. Foster et al. (2001) introduce a consistent semi-parametric estimation method without assuming parametric assumptions on the error distribution. In general, these robust adaptations are not suitable for heavy contamination and heteroscedasticity. Therefore, Marazzi and Yohai (2006) derive a consistent estimation method based on the minimization of a robust measure of residual autocorrelation with respect to a robust fit of the transformed outcome variable. This approach is robust to outliers, even if normality and homoscedasticity are not present in the data (Marazzi and Yohai, 2004).

C.1: A robustified maximum likelihood estimator

Carroll (1980) develop a more robust version of the profile log-likelihood estimator motivated by a dilemma. On the one hand, as shown by Andrews (1971), the normal maximum likelihood is usually not robust to deviations from normality or outliers. Andrews (1971) proposes a more robust method to overcome the sensitivity to outliers of the likelihood methodology based on the F -test of significance. On the other hand, Atkinson (1973) shows in a Monte Carlo experiment that the original likelihood test proposed in Box and Cox (1964) is more powerful than the significance method introduced by Andrews (1971). Atkinson (1973) suggests a modified version of the ML approach that does not account for robustness. This leads to the situation where a powerful method delivers no robust results, while a more robust method seems not to be so powerful. Based on the Huber's method (Huber, 1992) and the profile log-likelihood methodology presented earlier, Carroll (1980) propose an estimator which considers not only the normal distribution but also distributions with "normal-centre" and "exponential-tails". The method is powerful for these types of distributions, but also relatively robust to Andrew's method of significance. The likelihood function for such distributions is given by

$$L(\lambda, \beta, \sigma_e^2) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}} \sum_{i=1}^n \exp \left\{ -\rho \left(\frac{y_i^*(\lambda) - \mathbf{x}_i^T \beta}{2\sigma_e^2} \right) + (1 - \lambda) \log y_i \right\}, \quad (5.5)$$

where for some k and variable z

$$\rho(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq k; \\ k(|z| - \frac{k}{2}) & \text{if } |z| > k = 0. \end{cases}$$

Typical values of k are 1.5 or 2 (Carroll, 1980). Note that if $k = \infty$, Equation 5.5 is the normal likelihood for the Box-Cox transformation. λ , σ_e^2 and β are found in several stages. For further description of this algorithm please refer to Carroll (1980).

D: Bayesian approaches

As mentioned before, the ML estimator is not consistent in case non-normal errors are present and it is a non-robust methodology in the presence of outliers. Therefore, some research effort has been shifted towards alternative Bayesian estimation methods of the transformation parameter. The paper of Box and Cox (1964) propose a Bayesian estimation method for the transformation parameter, which uses a non-informative prior distribution of λ but is outcome-dependent. Pericchi (1981) introduced a solution for choosing a non-outcome-dependent a priori distribution, with a posterior log likelihood distribution similar in concept to the profile log likelihood-based on ML theory. Additionally, Sweeting (1984) suggested the use of a non-outcome-dependent family of non-informative priors distributions, which is closer in concept to that proposed by Box and Cox (1964).

5.2.2.2 Transformations to achieve homoscedasticity

Why is the homoscedasticity assumption important?

In linear regression analysis, the level of variance is assumed to be constant across the range of explanatory variables. This so-called homoscedasticity of the error term in the linear model can be formally written as $V(e_i|x_{i0}, \dots, x_{ip}) = \sigma_e^2$. It means that the conditional variance of e_i , given the set of values x_{i0}, \dots, x_{ip} , is not dependent on the x s (Wilcox, 2005). On the other hand, when the contrary occurs, we talk about heteroscedastic error terms, which can be expressed as $V(e_i|x_{i1}, \dots, x_{ip}) = \sigma_{e_i}^2$. What happens when the assumption of homoscedasticity is violated? As stated in many econometrics textbooks, the ordinary least squares (OLS) estimators for the β s remain unbiased and consistent but are no longer efficient or best linear unbiased estimator (BLUE) (Williams et al., 2013). This means, the OLS estimator does not provide the smallest variance or the smallest standard error estimations (see Wooldridge, 2000). Therefore, if the interest lies only in the estimation of the β s, OLS can be used. However, if the focus is on inference, then t -tests, F -tests, and confidence intervals are no longer valid since there is a higher probability of y lying outside the confidence interval, for example, for large values of x . Heteroscedasticity can arise from different sources: first, as a result of a measurement error, for instance coming from the fact that some respondents give more precise answers (Berry, 1993); second, from misspecifications of the model, e.g., when an important variable is omitted and thus, the error term exhibits idiosyncratic variation (Wooldridge, 2000); third, when the population should be clustered and thus variance changes across subpopulations (Natrella, 2013); and fourth, if there are outliers which means one or a few observations severely affect the non-robust variance estimator and induce (apparent) heteroscedasticity (Carroll, 1980). Some papers related to the consequences when homoscedasticity assumptions are not satisfied are in Cochran (1947) and Eisenhart (1947).

How can we check the homoscedasticity assumption fulfillment?

To graphically explore the homoscedastic assumption, let us suppose that we want to regress y against a vector containing one single explanatory variable, x . If the error term is homoscedastic, we would expect the set of points $[x, y]$ to spread along the regression line on the scatterplot exhibiting the same level of variation. A visual inspection of heteroscedasticity is made by plotting the residuals against the fitted values and the residuals versus a predictor which is possibly generating the violation of this assumption. There is also a huge range of tests for assessing homoscedasticity in the literature (Kirk, 1968). For detecting any linear form of heteroscedasticity, the Glesjer, Breusch-Pagan, Goldfeld-Quandt and Cook-Weisberg tests are commonly used. Additionally, the White's general test is useful when non-linear forms of heteroscedasticity need to be proved. Other suitable tests are the Hartley's Fmax (Hartley, 1950) and Cochran's C (Cochran, 1941), but they are sensitive to Gaussian assumptions, and the Bartlett's test (Bartlett, 1937) and Levene's test (Levene, 1960), among others. Additionally, the Ramsey Regression Equation Specification Error Test (RESET) test (Ramsey, 1969) can be used for the misspecification of the model.

What are the alternative methods to overcome heteroscedasticity?

If we have tested the correctness of the assumption and found statistical support to believe that the error term is heteroscedastic, a pre-analysis should be carried out before jumping into methods to correct for heteroscedasticity. First, model misspecification should be left to field experts for methodological issues. This is because heteroscedasticity arising from model misspecification is not genuine heteroscedasticity, but model misspecification since, by re-specifying the model, one could get rid of it. The need of clustering should be examined as well. It is also recommended to remove or replace outliers, or just apply an outlier treatment and then test for heteroscedasticity to verify if the homoscedasticity assumption is being violated by the influence of one or a few observations.

Alternatively, if the error term exhibits heteroscedasticity, a more robust and efficient estimator can be achieved via modified OLS residuals or generalized least squares (see Wooldridge, 2000). It includes the use of feasible generalized least squares and weighted least squares regression by minimizing a weighted sum of squared residuals (Berry, 1993). The downside of the latter is that the form of the weights is often unknown. Secondly, techniques for estimating robust standard errors can be used. They are known as heteroscedasticity-consistent-, Huber-, Eicker-, White-, Eicker-Huber-White-, Huber-White-standard errors or sandwich estimators (Eicker, 1967; Huber, 1967; White, 1980). Thirdly and most widely used, is the application of generalized linear regression models (Nelder and Wedderburn, 1972). These models take specific heteroscedasticity forms into account and contain different data structures; for instance, logistic regression for dichotomous (binary) variables or Poisson regression for count data. Additionally, Bayesian linear regression approaches can also account for the lack of homoscedasticity.

How can transformations help to improve homoscedasticity?

According to Johnson (1949) and based on Bartlett (1937) and Bartlett (1947), transformations might provide a fair correction for heteroscedasticity. When a functional dependence of the variance of the outcome variable on the mean is present in the data, we may gain the advantages of using variance-stabilizing transformations. This dependence mostly implies an underlying distributional process and determines the form of the suitable transformation. Table 5.2 shows different relations between these moments, the corresponding suitable transformations, some examples of appropriate distributions and the range of the outcome variable that the transformation supports. According to Ruppert (2001), populations which have larger means also exhibit the property of larger variances. If we denote the mean of the conditional distribution of the outcome variable given a vector of explanatory variables by $E(y|x) = \mu_y(x)$, then it is possible that the conditional variance $\text{Var}(y|x)$ is a function of $\mu_y(x)$. This relation is denoted by $\text{Var}(y|x) = R[\mu_y(x)]$ for some function $R(\cdot)$. Without loss of generality and following Ruppert (2001), if we use a transformation $T(y)$, this relation holds the delta-method linear approximation and is denoted as follows (Bartlett, 1947):

$$\text{Var}[T(y)|x] \approx \left\{ T'[\mu_y(x)] \right\}^2 R[\mu_y(x)].$$

The transformation $T(y)$ will correct the variance assumptions, if $[T'(y)]^2 R(y)$ is constant. For instance and following Ruppert (2001), if $R(y) \propto y^\alpha$, then $T(y) \propto y^{1-\frac{\alpha}{2}}$ would be a variance-stabilizing transformation, with $\alpha \neq 2$. The transformations that are used most for the issue of achieving homoscedasticity are square roots, logarithms, reciprocals, and trigonometrical transformations (Cohen et al., 2014). Some of these are also known as double bend transformations, because the data sets for which these are used, are bound at both top and bottom, such as data sets from the binomial distribution.

Bartlett (1937) proposes the use of the square root transformation to stabilize variances that are exactly proportional to the mean, which is the case for gamma and exponential distributed data, as for such a distribution in which the variance is exactly equal to the mean, which is the case of the Poisson distribution. In this case, $\alpha = 1$, that means, $g(y) \propto y^1$, then $T(y) \propto y^{1-\frac{1}{2}}$ is the square root, which is the variance-stabilizing transformation for Poisson data sets. In a later work, Bartlett (1947) and Anscombe (1948) suggest the use of $\sqrt{y + c_1}$ type transformations, where c_1 is a fixed constant. In case of a large sample size this transformation with a constant c_1 is more useful to achieve a constant variance. They propose handling heteroscedasticity by using $c_1 = 1/2$ or $c_1 = 3/8$, when y takes only small values or when zeros are common in the data, respectively. Freeman and Tukey (1950) proposes a more sophisticated twofold transformation, which is called the Freeman-Tukey deviate or the chordal transformation and is denoted by $\sqrt{y} + \sqrt{y+1}$. This transformation is particularly suitable in case y is very small or equal to 0. Similarly, the inverse transformation is recommended for stabilizing the variance for observations that are mostly close to zero. It stabilizes the variance when $n > 3$ (Mosteller and Bush, 1954; Mosteller and Youtz, 2006).

The negative binomial distribution is appropriate to represent for Poisson distributed data under overdispersion, that means, the variance greater than the mean. For this kind of data sets

Table 5.2: Transformations for achieving homoscedasticity.

Dependence	Source	Formula	Example	Support
$\sigma_y^2 \propto \mu_y$	Bartlett (1937)	\sqrt{y}	Poisson(λ)	$y \geq 0$
$\sigma_y^2 \propto \mu_y$	Bartlett (1947)	$\sqrt{y + c_1}$	Poisson(λ)	$y \geq -c$
$\sigma_y^2 \propto \mu_y$	Freeman and Tukey (1950)	$\sqrt{y} + \sqrt{y + 1}$	Poisson(λ)	$y \geq -1$
$\sigma_y^2 \propto \mu_y^2$	Fisher and Yates (1949)	$\log(y)$	lognormal(μ, σ^2)	$y > 0$
$\sigma_y^2 \propto \mu_y^2$	Fisher and Yates (1949)	$\log_{10}(y)$	lognormal(μ, σ^2)	$y > 0$
$\sigma_y^2 \propto \mu_y^2$	Fisher and Yates (1949)	$\frac{1}{3}\sqrt{y}$	lognormal(μ, σ^2)	$y \geq 0$
$\sigma_y^2 \propto 2\mu_y$	Freeman and Tukey (1950)	$\sqrt{2y}$	$\chi^2(k)$	$y \geq 0$
$\sigma_y^2 \propto 2\mu_y$	Wilson and Hilferty (1931)	$y^{1/3}$	$\chi^2(k)$	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu_y + \lambda^2 \mu_y^2$	Bartlett (1947)	$\lambda \log(y)$	BN($r, p, \lambda = \frac{1}{\sqrt{r}}$)	$y > 0$
$\sigma_y^2 \propto \mu_y + \lambda^2 \mu_y^2$	Bartlett (1947)	$\log(y)$	BN($r, p, \lambda = \frac{1}{\sqrt{r}}$)	$y > 0$
$\sigma_y^2 \propto \mu_y + \lambda^2 \mu_y^2$	Bartlett (1947)	$\lambda^{-1} \sinh^{-1}(\lambda\sqrt{y})$	BN($r, p, \lambda = \frac{1}{\sqrt{r}}$)	$y \geq 0$
$\sigma_y^2 \propto \mu_y + \lambda^2 \mu_y^2$	Bartlett (1947)	$\lambda^{-1} \sinh^{-1}\left(\lambda\sqrt{y + \frac{1}{2}}\right)$	BN($r, p, \lambda = \frac{1}{\sqrt{r}}$)	$y \geq -\frac{1}{2}$
$\sigma_y^2 \propto \mu_y + \lambda^2 \mu_y^2$	Beall (1942)	$\sinh^{-1} \sqrt{\frac{y+c_2}{r+c_3}}$	BN($r, p, \lambda = \frac{1}{\sqrt{r}}$)	$y \geq 0$
$\sigma_y^2 \propto \mu_y$	Ruppert (2001)	$\log(y)$	BN($r, p, \lambda = \frac{1}{\sqrt{r}}$)	$y > 0$
$\sigma_y^2 \propto \mu_y$	Watthanacheewakul (2014)	$\begin{cases} \frac{(\sqrt{y+1})^\lambda}{\lambda} & \text{if } \lambda \neq 0; \\ \log(\sqrt{y+1}) & \text{if } \lambda = 0. \end{cases}$	$\Gamma(\alpha, \beta)$, Weibull(l, k)	$y \geq -1$
$\sigma_y^2 \propto \mu_y$	Wilson and Hilferty (1931)	$y^{1/3}$	$\Gamma(\alpha, \beta)$	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu_y$	Curtiss (1943)	$\sqrt{y + c_1}$	$\Gamma(\alpha, \beta)$	$y \geq -c$
$\sigma_y^2 \propto \mu_y$	Ruppert (2001)	$\log(y)$	$\exp(\lambda)$	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Bartlett (1937)	$\sin^{-1} \sqrt{y}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Bartlett (1937)	$\sin^{-1} \sqrt{\frac{y+c_4}{n+c_5}}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Anscombe (1948)	$\sqrt{n + c_6} \sin^{-1} \sqrt{\frac{y+c_4}{n+c_5}}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Laubscher (1961)	$\sqrt{n} \sin^{-1} \sqrt{\frac{y}{n} + \sqrt{n+1} \sin^{-1} \sqrt{\frac{y+\frac{3}{4}}{n+\frac{3}{2}}}}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Freeman and Tukey (1950)	$\sqrt{n + \frac{1}{2}} \left(\sin^{-1} \sqrt{\frac{y}{n+1}} + \sin^{-1} \sqrt{\frac{y+1}{n+1}} \right)$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Fisher (1922b)	$\sin^{-1} y$	Bin(n, p)	$0 \leq y \leq 1$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Fisher (1922b)	$\sin^{-1} \sqrt{\frac{y+c_4}{n+c_5}}$	Bin(n, p)	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Curtiss (1943)	$\sqrt{n} \sin^{-1} \sqrt{y + \frac{c_7}{n}}$	Bin(n, p)	$y > 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Curtiss (1943)	$\sqrt{n} \log(y)$	Bin(n, p)	$y > 0$
$\sigma_y^2 \propto \mu_y(1 - \mu_y)$	Curtiss (1943)	$\frac{1}{2} \sqrt{n} \log\left(\frac{y}{1-y}\right)$	Bin(n, p)	$y > 0$
$\sigma_y^2 \propto \mu_y^3$	Draper and John (1981)	$\frac{1}{\sqrt{y}}$	-	$y > 0$
$\sigma_y^2 \propto \frac{1}{\mu}$	Draper and John (1981)	y^2	-	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu^2$	Draper and John (1981)	$\log(y)$	-	$y > 0$
$\sigma_y^2 \propto \mu_y$	Draper and John (1981)	\sqrt{y}	-	$y \geq 0$
$\sigma_y^2 \propto \mu_y^4$	Draper and John (1981)	$\frac{1}{y}$	-	$y \neq 0$

Note: Please note that due to lack of different parameter names and conventional definitions of the distributions in column Example the parameter names can conflict with the notation in the rest of the paper. The parameter $c_1 = 1$ is widely used in practice. However, Bartlett (1937) and Anscombe (1948) recommend $c_1 = \frac{1}{2}$ and $c_1 = \frac{1}{3}$, respectively. Beall (1942) suggests $c_2 = c_3 = 0$ and Anscombe (1948) $c_2 = \frac{3}{8}$, $c_3 = \frac{-3}{4}$. This author recommends $c_4 = \frac{3}{8}$ and $c_5 = \frac{3}{4}$, meanwhile Bartlett (1937) $c_4 = \frac{1}{2}$ and $c_5 = 0$. However, $c_4 = c_5 = 0$ are often used in practice. Anscombe (1948) suggests $c_6 = \frac{1}{2}$. Curtiss (1943) suggests c_7 equal to 0 or $\frac{1}{2}$, depending on the values of p .

some transformations based on the logarithm and hyperbolic trigonometric functions are proposed (Bartlett, 1947; Chatterjee and Hadi, 2015). For instance, some modifications of the inverse hyperbolic sine function, such as $\sinh^{-1} \sqrt{\frac{y+c_2}{k+c_3}}$, are suitable for the negative binomial data. While Anscombe (1948) suggests values of $c_2 = 3/8$ and $c_3 = -3/4$, Beall (1942) proposes using $c_2 = 0$ and $c_3 = 0$. Especially recommended for small values is the adjustment $\frac{1}{\lambda} \sinh^{-1}(\lambda \sqrt{y + 1/2})$ (Chatterjee and Hadi, 2015).

In order to stabilize the variance of binomial distributed data, different trigonometric transformations are suggested. For instance, the inverse sine root square transformation of the form $\sin^{-1} \sqrt{y}$, also called the angular transformation (Fisher, 1922b; Bartlett, 1937), is analogous to the root square transformations for binary data. Thus, this variance-stabilizing transformation is widely used in practice. Some modifications based in this transformation have been proposed according to specific values of the parameters of the distribution based on the data set are proposed in Curtiss (1943) and Anscombe (1948). For instance, $\sin^{-1} \sqrt{\frac{y+c_4}{n+c_5}}$ is then suitable for data from the binomial distribution. Researchers commonly use $c_4 = c_5 = 0$. However Bartlett (1937) suggests $c_4 = 1/2$ and $c_5 = 0$ and Anscombe (1948) improves this transformation further by setting $c_4 = 3/8$ and $c_5 = 3/4$. Unlike similar transformations, the arcsine is defined for y between 0 and 1. However, research done by Wilson et al. (2013) and Warton and Hui (2011) have warned about employing this transformation. According to Warton and Hui (2011) one of the downsides of this transformation is that if the relation between the untransformed y and the independent variables x_{i0}, \dots, x_{ip} is e.g., always increasing, the same relation is not held after transformation due to the periodicity of arcsin. Sophisticated twofold transformations are also suggested by Laubscher (1961) and Freeman and Tukey (1950). To correct for heteroscedasticity of variables contained to a bounded interval, such as proportions and percentages, two-bend transformations families can be appropriate. For instance, the most common transformations are the logit, probit, Guerrero-Johnson, Aranda-Oraz, beta, angular and arsine transformations. For detailed information about transformations for these kinds of data sets please refer to Kruskal (1968); Atkinson (1987) and Piepho and McCulloch (2004). This topic falls out of the scope of the present work.

The ordinary power transformations family, in which different powers of the target variable are applied, are defined according to the functional dependence of the variance on the mean. If the variance increases proportional to the mean on a square root scale, the stabilization is made on a logarithmic scale (Bartlett, 1947). This is the case of the log-normal distribution. Fisher and Yates (1949) proposed some modifications of the logarithmic transformation in case the values are less than 10 and for larger values. For distributions with constant coefficient of variation, such the exponential or gamma with constant shape parameter distribution, the logarithm is also recommended (Ruppert, 2001). This transformation is generally suggested when the range of the outcome variable is very broad but not negative (Fink, 2009). For data from other distributions as the Gamma or Weibull distribution a variance stabilizing transformation is recently proposed by Watthanacheewakul (2014). When data is very bunched to the minimum and maximum of the distribution the transformation presented by Fink (2009) can be used for stretching the data. For selecting the parameters λ and k of this transformation we refer to Fink (2009); Erickson and Nosanchuk (1977) and McNeil (1977). Additionally, if the data

presents heteroscedasticity problems and the distributional form is not clear or there are other violations of assumptions, some of the already mentioned transformations in the beginning of Section 5.2.2 also help to correct heteroscedasticity, since stabilizing variance and normalizing errors often goes together (Johnson, 1949). These transformations include in particular the logarithm, gpower, Box-Cox, Johnson, Manly, and Yeo-Johnson transformations. That means, the researcher should empirically find the most appropriate transformation that stabilizes the variance of the data regardless the mean value (Montgomery, 2008). Finally, transforming both sides helps for both, stabilizing the variance and create more symmetric distributions (see Section 5.3 for the both sides methodology).

How can we estimate the transformation parameters to homoscedasticity?

In general, the approaches for estimating the optimal transformation parameter to homoscedasticity are ML-based or analytical considerations. Therefore, in case a transformation for simultaneous correcting non-normality and heteroscedasticity is selected, then the ML-based approaches presented already for normality can be used. However, Zarembka (1974b) pointed out that this method is not robust in the presence of heteroscedastic error terms. Therefore, Blaylock and Smallwood (1985) propose an alternative adaptation, the robustified maximum likelihood estimator. Hinkley (1985) suggests the use of an analytical likelihood-based method for analyzing local deviations. This procedure for estimating the transformation parameter considers both the homoscedasticity model violation of residuals and the lack of additivity. Ruppert and Aldershof (1989) propose a method which attempts to deal with non-normality and heteroscedasticity. It is based on the minimization of the correlation between the fitted values and the squared residuals.

5.2.2.3 Transformations to achieve linearity and additivity

Why is the linearity assumption important?

As it is implied in its name, the linear regression is an approach to model linear relationships. The linear regression model is linear in two senses: first, the model is linear in the variables because each response y is expressed as a weighted sum of the independent variables where the parameters are the weights (Dougherty, 2011); second, the model is also linear in the parameters where, this time, the independent variables are the weights. If non-linearity is present and we decide to follow through with the use of linear techniques as in OLS, the consequences would be misrepresenting the actual relationship. Therefore, when non-linearity occurs, it is very likely that estimation and inference techniques based on the linearity of the model yield misleading conclusions. In addition to linearity, it is important that the additivity assumption is met. This assumption ensures that the independent variables multiplied by their regressors can be added together to provide an estimate (Berry, 1993). However, given the complexity of many empirical relationships, it is sometimes expected that the effect of an independent variable x_1 on y may be influenced by a third variable, x_2 . This interaction not only violates the implicit assumption of additivity, but it also becomes a practical problem since it leads to multicollinearity (Friedrich, 1982). Moreover, when a non-additive relationship takes place, and it

is not detected or is ignored, the linear regression yields unreliable results since the relationship that is being represented fails to account for the interaction between the independent variables. Again, as in the presence of non-linearity, estimation and inference techniques based on the linear regression model provide non-accurate results (Williams et al., 2013).

How can we check the linearity assumption fulfillment?

A useful visual method to examine non-linearity is using scatterplots between the outcome variable and the explanatory variables, which is called added variable plot, also known as partial-regression- or adjusted plot (Atkinson, 1982). Additionally, a scatter plot of the standardized residuals and the standardized predicted values of y is also useful. If the relationship appears to take a line-like form, we do not need to occupy ourselves with correcting for non-linearity. Additionally, the RESET test, a general test for functional form misspecification proposed by Ramsey (1969, 1974) can be used as an indicator of lack of linearity.

A technique to detect non-additivity effects is the Tukey's test (Tukey, 1949; Moore and Tukey, 1954). As an alternative to Tukey's test, Barry (1993) introduces a Bayesian test to check the validity of this assumption.

What are the alternative methods to overcome non-linearity?

If the assessment tools provide evidence for non-linearity and/or non-additivity, a model restructuring is a possible solution. For instance, if the relation between the dependent and independent variables seems to be curvilinear, a curve component could be added and tested on significance (Osborne and Waters, 2012). For receiving additivity, Friedrich (1982) favors the use of multiplicative models over dropping interactive variables to use linear regression techniques. If non-linearity or non-additivity is still present, ridge regression, also known as linear regularization, is particularly useful. Other alternative methods are Tikhonov regularization, Tikhonov-Miller method, Phillips-Twomey method, constrained linear inversion method or weight decay (Hoerl and Kennard, 1970), lasso regression (Tibshirani, 1996) and Bayesian linear regression.

How can transformations help to improve linearity?

In general, transformations to linearize data can be divided into two classes: in one class, the expected response is related to the independent variables by a known non-linear function; in the other, the relationship between the expected response and the explanatory variables is not exactly known (Cook and Weisberg, 1982). For the first class, transformations can be easily selected. Cuthbert et al. (1971) show plots for a comprehensive number of non-linear functions that can be transformed into linear ones. In the second class fall transformations such as the Box-Cox transformation, which have the potential to correct non-normality, heteroscedasticity, and non-linearity, so that, after the data is transformed, normal theory methods and linear regression techniques can be employed. An approach for selecting a suitable power transformation is given by Mosteller and Tukey (1977), who introduce a trial-and-error heuristic to linearize data based on the ladder of powers, called the "ladder of transformations", as shown

Table 5.3: Transformations to achieve linearity when the relation is known for the simple linear regression model.

Reference	Regression form	Transformation	Linear model
Weisberg (1980)	$y = \beta_0 x_1^\beta$	$y^* = \log y, x^* = \log x$	$y^* = \log \beta_0 + \beta_1 x^*$
Weisberg (1980)	$y = \beta_0 e^{\beta_1 x}$	$y^* = \log y$	$y^* = \log \beta_0 + \beta_1 x$
Weisberg (1980)	$y = \beta_0 + \beta_1 \log x$	$x^* = \log x$	$y^* = \beta_0 + \beta_1 x^*$
Weisberg (1980)	$y = \frac{x}{\beta_0 x - \beta_1}$	$y^* = \frac{1}{y}, x^* = \frac{1}{x}$	$y^* = \beta_0 - \beta_1 x^*$
Chatterjee and Hadi (2015)	$y = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$	$y^* = \log \left(\frac{y}{1-y} \right)$	$y^* = \beta_0 + \beta_1 x$
Fink (2009)	$y = \beta_0 + \beta_1 \left(\frac{1}{x} \right)$	$x^* = \frac{1}{x}$	$y^* = \beta_0 + \beta_1 x^*$
Weisberg (1980)	$y = \frac{1}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}$	$y^* = \beta_0 + \beta_1 x$
Johnson (2009)	$y = \beta_0 + \beta_1 \sqrt{x}$	$x^* = \sqrt{x}$	$y^* = \beta_0 + \beta_1 x^*$

in Table 5.4 and Table 5.5. This is also known as the “bulging rule” and it determines the value of the power employed for both the outcome variable and the explanatory variables within the model. Please follow Brown (2015) for more details about the bulging rule. Power transformations are useful if the relationship between x and y is a simple monotone. Table 5.3 summarizes transformations for regression forms that can be linearized since the relation is known for the simple linear regression model, and is based on Weisberg (1980); Fink (2009); Johnson (2009) and Chatterjee and Hadi (2015). The generalization of this table for the multiple regression form can be find in Fink (2009). Additionally, Box and Tidwell (1962) propose an iterative methodology known as the Box-Tidwell transformation to linearize the relationship between the dependent variable and the explanatory variables. It is basically based on individually finding the optimal power transformation to transform the set of explanatory variables. A power transformation test can help to determine which variable should be transformed or not (Brown, 2015). Finally, both sides methodology is also suitable for dealing with non-linearity problems in the regression model (see Section 5.3). Nevertheless, one should be careful when transforming both sides to induce linearization, since it may produce heteroscedasticity of the error term (Carroll and Ruppert, 1988). A tentative transformation to linearize multiplicative models is the logarithmic transformation. For non-additivity, Tukey (1949) recommends the use of the t -score of added non-linear terms as the transformation criteria. Without loss of generality, the transformations that are suitable for correcting non-additivity have a restricted form and the works of Elston (1961) and Anscombe and Tukey (1963) concentrate on the selection of the power. Rocke (1993) suggests the use of the t -score as a criteria to linearize proportional data.

How can we estimate the transformation to linearity?

For the transformations that fall in the second class, the ML-based methods and analytical considerations that we already introduced are equally applicable for achieving linearity. A special approach to find the correct power when the regression form is known is given by Mosteller and Tukey (1977), who introduce a trial-and-error heuristic to linearize data based on the ladder of powers shown in Table 5.4 and Table 5.5. Tukey (1949) introduces the minimization of the F -value for the degree of freedom for non-additivity as an estimation method of a transforma-

Table 5.4: The ladder of powers.

λ_i	-2	-1	-0.5	0	0.5	1	2
z	$\frac{1}{z^2}$	$\frac{1}{z}$	$\frac{1}{\sqrt{z}}$	$\log z$	\sqrt{z}	z	z^2

tion.

The Tukey and Mosteller estimation algorithm

As mentioned before, Mosteller and Tukey (1977) propose a graphical bulging rule for selecting a power transformation, which is based on power of ladders. This seeks to guide practitioners to simply select a linearizing relationship transformation. For any random variable z , the ladders are tabulated as follows: They can be generalized and formally expressed as:

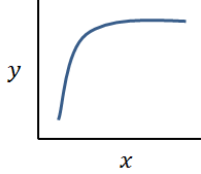
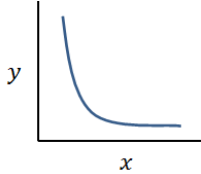
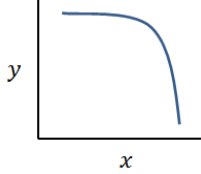
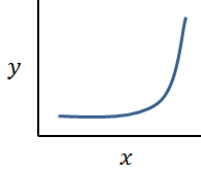
$$y_i^*(\lambda) = \begin{cases} y_i^{\lambda_1} = \beta_0 + \beta_1 x_i^{\lambda_2} & \text{if } \lambda_1, \lambda_2 \neq 0; \\ \log yi = \beta_0 + \beta_1 x_i^{\lambda_2} & \text{if } \lambda_1 = 0, \lambda_2 \neq 0; \\ y_i^{\lambda_1} = \beta_0 + \beta_1 \log x_i & \text{if } \lambda_1 \neq 0, \lambda_2 = 0; \\ \log yi = \beta_0 + \beta_1 \log x_i & \text{if } \lambda_1, \lambda_2 = 0. \end{cases}$$

The parameters λ_1 and λ_2 are chosen according to Table 5.4 and Figure 5.5. Examining a scatterplot of y against x leads us to select a power transformation based on the pattern of the curvature. We have two options for transforming: transform y by moving up/down the ladder or up/down the ladder for x depending on the pattern. That means in case the pattern is hollow upward, one should go down the ladder; and if hollow downward go up the ladder.

Mosteller and Tukey (1977) present a simple numerical algorithm, which is explained as follows:

1. Plot x against y .
2. Based on Table 5.4, choose λ_1 and λ_2 according to the shape exhibited by the points on the scatter plot of x against y .
3. Transform y by y^{λ_1} and x by x^{λ_2} .
4. Plot the transformed predictor against the transformed response variable.
5. If the relationship appears to be linear: stop.
6. Otherwise, choose new values for λ_1 and/or λ_2 by going up or down the power ladder based on Table 5.4.

Table 5.5: The ladder of transformations.

Pattern	Transformation	Parameter
	$y^* = y^{\lambda_1 > 1}, x^* = x^{\lambda_2 < 1}$	λ_1 up and/or λ_2 down
	$y^* = y^{\lambda_1 < 1}, x^* = x^{\lambda_2 < 1}$	λ_1 down and/or λ_2 down
	$y^* = y^{\lambda_1 > 1}, x^* = x^{\lambda_2 > 1}$	λ_1 up and/or λ_2 up
	$y^* = y^{\lambda_1 < 1}, x^* = x^{\lambda_2 > 1}$	λ_1 down and/or λ_2 up

5.2.3 Parameter inference and interpretation

Does a transformation influence the inference on the model parameters?

The inference analysis is a controversial question that arises when a transformation, and especially a transformation with a transformation parameter, is used under the linear and linear mixed regression model. One question is whether we should treat the transformation parameters as fixed in case we are making inferences on the model parameters. If the transformation does not contain a data-driven transformation parameter common model inference can be conducted. In contrast, when using data-driven transformations, one point of discussion concerns if the transformation parameter can be treated as known or not. Ruppert (2001) and Box and Cox (1982) stated that the regression parameter estimates strongly depend on the chosen transformation parameter λ . Box and Cox (1964) further pointed out that after selecting a value for λ via e.g., ML-based methods, this should be treated as known and inference can be carried out as usual. However, Bickel and Doksum (1981) made a remark on this by studying the joint distribution of $\hat{\lambda}$ and $\hat{\beta}$. They found that when the real value of λ is unknown, the estimates for the variance of the $\hat{\beta}$ s are inflated and highly dependent on the $\hat{\lambda}$ estimate. Box and Cox (1982, p. 209) replied by saying that this was not only obvious, but also irrelevant, since “the gross correlation effects would be avoided if, following [their] paper, the investigation had been conducted in terms of [the normalized transformation]”. Note that the normalized transformation is equivalent to the scaled transformation presented in Section 5.2.2. Furthermore, Hinkley and Runger (1984) carried out a sensitivity analysis where they found that the estimates of contrast

and scale parameters are quite stable on the scale of the normalized transformation, whereas the estimates of location parameters, such as the mean, are more dependent on the value of $\hat{\lambda}$.

Research on the accuracy of the estimation and inference on the random effects after applying a transformation under a linear mixed regression model is still necessary. Under this scenario, the works of Verbeke and Lesaffre (1996) and Gurka et al. (2006) discussed, in a simulated scenario, the effects of a transformation on the inference process. Gurka et al. (2006) suggests including a correction factor from the Jacobian of the Box-Cox transformation in the estimated coefficients.

How is the inference process on the transformation parameters?

Inference about the transformation parameters is also a fundamental step in the transformation selection process. For testing the hypothesis $H_0 : \lambda = \lambda_0$, we could use the standard likelihood-based methods for getting a likelihood ratio test. The test statistic would be $W = 2[L_{\max}(\hat{\lambda}) - L_{\max}(\lambda)]$, which is asymptotically chi-squared distributed. Box and Cox (1964) extend this theory and propose two approaches to make inferences about the parameters after applying a transformation. In the first approach, large sample maximum likelihood theory is applied, which delivers point estimates of the parameters and provides an approximate test and confidence intervals based on the chi-squared distribution. In the second approach, Bayesian theory is applied. For that, the prior distributions for β and σ^2 are assumed to be uniform, obtaining a posterior distribution for λ . For more details about the Bayesian method please see Box and Cox (1964) and Jeffreys (1998).

Following Box and Cox (1964), an approximate $100(1 - \alpha)$ per cent confidence interval is

$$L_{\max}(\hat{\lambda}) - L_{\max}(\lambda) < \frac{1}{2}\chi_{\nu}^2(\alpha),$$

$$L_{\max}(\lambda) = -\frac{1}{2}n \log [\hat{\sigma}^2(\lambda)] + \log [J(\lambda, y)],$$

where ν is the number of independent components in λ , α denotes the significance level, and $\hat{\sigma}^2(\lambda)$ represents the residual sum of squares in the transformed outcome variable.

In the same way, in order to test $H_0 : \lambda = \lambda_0$, Andrews (1971) proposes a test which ignores the Jacobian of the applied transformation. However, Atkinson (1973) re-introduces the Jacobian, developing a score-type statistic, which is not a maximum likelihood-based method. This is also known as the Atkinson's score statistic and was further standardized by Lawrance (1987a,b), the result of which is called Lawrance's statistic. Some robust versions of these tests are proposed by Carroll (1980) and Wang (1987).

Last but not least, some studies regarding the consistency and efficiency properties, as well as the asymptotic variances of the estimated λ in the Box-Cox transformation, have been published. See Bickel and Doksum (1981); Carroll and Ruppert (1981); Carroll (1982a); Doksum and Wong (1983) and Hinkley and Runger (1984) for detailed information.

How are the model results interpreted when a transformation is applied?

One of the biggest challenges that researchers face when working with transformations is the interpretation of the results. It implies choosing the scale in which we need to present the results, depending on the research question. O'Hara and Kotze (2010) summarized this issue by pointing out that transformations come at some cost to the trade-off between accuracy and interpretability. When working with the logarithmic transformation, an approximation helps to obtain a meaningful interpretation of the coefficients as percentages. However, this is a feature rarely observed when working with other non-linear transformations, such as the Box-Cox transformation family. In the words of Box and Cox (1964), transformation parameters that are obtained by maximum likelihood-based methods, which are widely used in practice for finding a suitable transformation, are "useful as a guide" but "not to be followed blindly". The selection of transformation parameters could be made based only on the information provided by the data. However, if a particular value for λ in the Box-Cox transformation is more convenient regarding interpretability, the selection of the parameter could be adjusted. For instance, if the output of an estimation suggests that λ should be equal to 0.25 one could work instead with $\lambda = 0$ i.e., the logarithmic transformation, which has an easier interpretation, especially when this choice is common in the specific research field.

Does the back-transforming process lead to a bias in the predictions?

Researchers interested in predictions face another challenge which is to deal with the back-transforming bias when applying non-linear transformations. In case, a back-transformation is used for getting the values in their original measurement scale, an artificial bias comes from this re-transforming process. Without loss of generality:

$$T[E(y|x)] \neq E[T(y)|x]$$

for all non-linear transformations, $T(\cdot)$ applied on the target variable. Although this effect is not always severe (Sakia, 1992), ignoring the magnitude of the generated bias may lead to misleading conclusions. Therefore, several methods and empirical work for removing the back-transforming bias after applying a power transformation, in particular, the logarithmic and Box-Cox transformations have been proposed in the literature for the linear and the linear mixed regression model (Neyman and Scott, 1960; Hoyle, 1973; Lee, 1982; Rothery, 1988; Sakia, 1990, 1992; Newman, 1993; Gurka et al., 2006; da Costa and Crepaldi, 2014).

5.3 Further issues with regards to variable transformations

Additionally to the model assumptions that we discussed in the previous section, special features in the data can interact with the transformations or have effects on the usage of transformations. Thus, this section discusses issues such as model selection, the presence of outliers, incomplete responses, multimodal data, zero inflated data, and the range of the variable when using transformations. Note that these issues are a selection of the most common possible interactions. Furthermore, this section explains how to decide which variables in the model

should be transformed.

How is the model selection process under transformations?

The strategy for selecting the working model under different transformation is still under discussion. Sakia (1992, p. 174) states “The selection of a transformation may be properly viewed as model selection”. However, comparing regression models for variable selection under different scale levels has some difficulties. The model selection criterion should be invariant to a change of scale in the target variable, which is not the case for the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), two commonly used information criteria for the linear and linear mixed regression models (Burnham and Anderson, 2004; Müller et al., 2013). Therefore, the coefficients of determination and their extensions to the linear mixed regression models are a first approximation for comparing the models in terms of general fitting, since they are scale invariant. Additionally, the working model always depends on which procedure is done first, variable or transformation selection. Some procedures that have been implemented for the linear regression model include the combination of these two procedures in one (Laud and Ibrahim, 1995; Hoeting and Ibrahim, 1998; Hoeting et al., 2002).

How should transformations be used in the presence of outliers?

Without loss of generality an outlier is defined as an atypical observation among a data set, which can be representative or non-representative (Chambers, 1986). A discussion of the definition of outlying observations for the linear mixed regression models can be found in Bell and Huang (2006) and Warnholz (2016b). Outliers are not themselves a violation of model assumptions. However, their presence could induce skewed distributions and heteroscedasticity that lead to problems already examined in Section 5.2. Simply excluding an outlier is not always the right answer since they may contain valuable information about the distribution of our data (Belsley et al., 2005). If the presence of an outlier has a disproportionate influence on the estimated model, an analysis with and without such observation is usually recommended.

Carroll and Ruppert (1985) state that proper transformation decision and identification of outlying observations are interconnected. Thus, Figure 5.1 summarizes the stages in the analysis where the detection and the handling of outliers interconnects with the usage of a transformation. For the detection of outliers, scatterplots between the outcome variable and the explanatory variables or a box plot of the outcome variable can be sufficient. More methodologies can be found for instance in Cook and Weisberg (1982) and Barnett and Lewis (1984). The most popular measures of influence are the Cook’s distance (Cook, 1977), the Welsch and Kuh measure (Belsley et al., 2005) and the Hadi’s Influence Measure (Hadi, 1992). In the case that the use of a transformation seems suitable after checking model assumptions and outliers are detected, a sensitivity analysis is suggested. This includes finding out if the outlying case in the original scale is also an outlying observation in the transformed scaled. Furthermore, it is important to have an idea about how these observations can influence the need or utility of a transformation. For instance, if the outliers cause heteroscedasticity, the deletion of the outlier could make the usage of the transformation unnecessary. Some diagnostics for studying the contribution of single observations on the need of transformations are presented in

Cheng (2005) and Atkinson and Riani (2012). A sensitivity analysis under a Box-Cox power transformation model has been discussed by Bickel and Doksum (1981); Box and Cox (1982); Hinkley and Runger (1984); Atkinson (1986) and Duan (1993). Atkinson (1986) proposes a sensitivity analysis by eliminating outlying observations after applying a Box-Cox transformation. Atkinson (1982) studied the reduction of influential cases and outliers after applying transformations in some examples. However, Cook and Wang (1983) proposed a method to detect influential observations under the Box-Cox transformation that is superior to the method of Atkinson (1982) (Cook and Wang, 1983; Sakia, 1992). Tsai and Wu (1990) and Kim et al. (1996) studied the influence on the Jacobian of the transformation when single observations are deleted. If the outliers influence the need of the transformation, different methods are suitable to treat the outliers (Hawkins, 1980; Cook and Prescott, 1981; Cook and Weisberg, 1982; Cook and Wang, 1983; Barnett and Lewis, 1984; Hawkins et al., 1984). In a model context and for the estimation process, different procedures have been proposed: model reformulations, downweighting outlying observations (Rousseeuw and Leroy, 2005), use of the winsorization method (Yale and Forsythe, 1976) and use extreme-value distributions (Withers and Nadarajah, 2007). Furthermore, there has been considerable growing interest in using robust techniques in recent years for incorporating this effect into the model structure and fitting or bounding outliers and influential observations (Huber, 1964; Krasker and Welsch, 1982; Hampel et al., 1986; Rousseeuw and Van Zomeren, 1990). For instance, the M-estimation (Huber, 1964) and the least trimmed squares (Anscombe and Guttman, 1960) are examples of robust models that can be used when outliers are present in the data. As alternatives, it is common in practice to use Bayesian methods (Gelman et al., 2014) and quantile regression (Koenker, 2005).

In the other case, transformations can be useful in the presence of outliers since all information can be kept in the data set and, at the same time, skewness and error variance can be reduced (Osborne and Overbay, 2004). Furthermore, the rank transformation (Conover and Iman, 1981) replaces the data for their corresponding ranks, and it can be seen as an outlying observations handling. When a transformation is used without a previous outliers treatment, it is recommended to use robust methods for the estimation of the transformation parameters because the maximum likelihood theory is sensitive to outliers. In particular, Carroll (1980) Carroll (1982b), Carroll and Ruppert (1985), Bickel and Doksum (1981) and most recently Marazzi and Yohai (2004) propose different robust methods for the Box-Cox transformation and Burbidge et al. (1988) for the inverse hyperbolic sine transformation parameters. These approaches are concerned with a modified likelihood function (see e.g., Krasker and Welsch, 1982). Gottardo and Raftery (2009) developed a Bayesian estimation method for the Box-Cox transformation that accounts for outlying values. Pericchi (1981) and Sweeting (1984) study different choices of prior distributions for the Box-Cox linear model. For the same model, Shin (2008) develops a semi-parametric estimation method. Note that all mentioned methods only handle outliers in the outcome variable.

How do incomplete responses affect the usage of transformations?

The problem of missing data becomes a fundamental part of almost every research setting. Rubin (1976) introduced a classification system of missing data which describes the probability

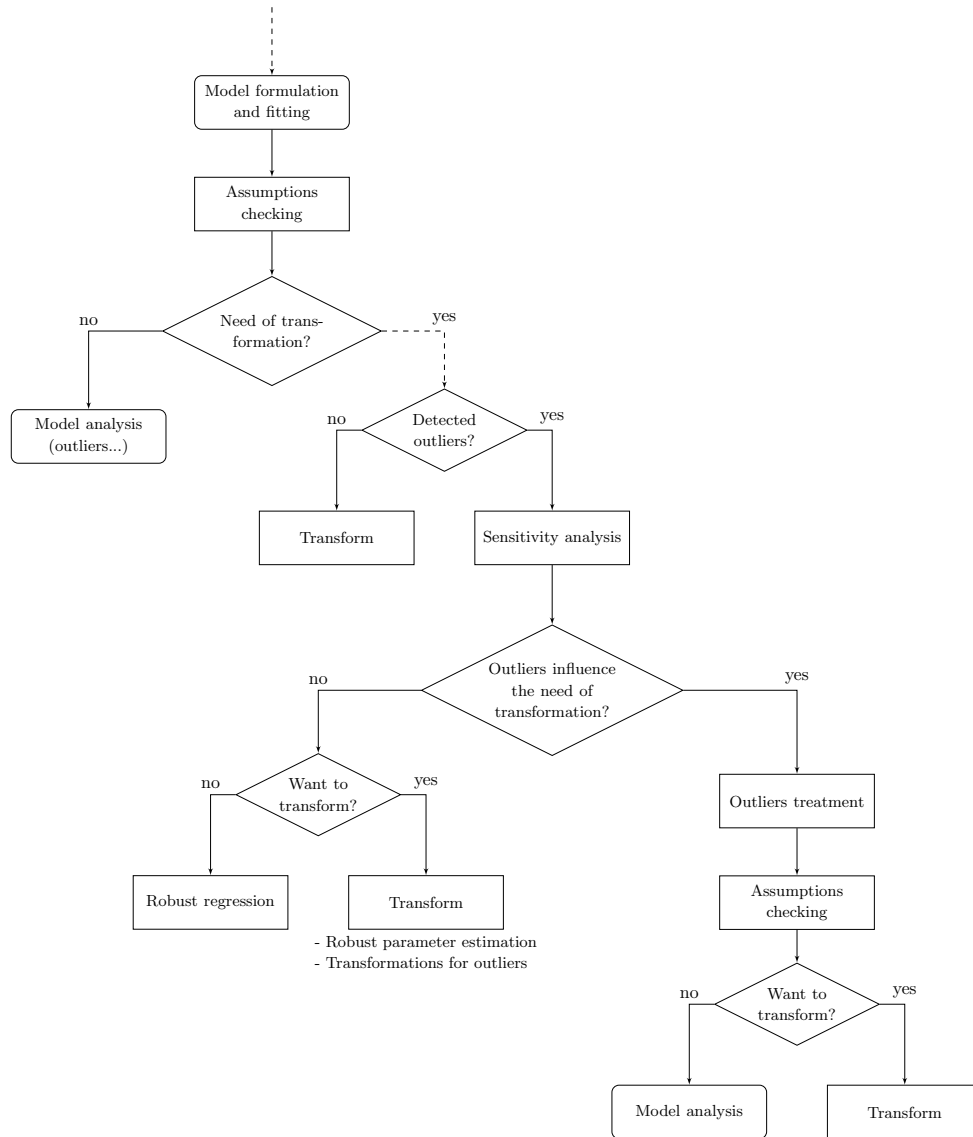


Figure 5.1: A guide how to handle the interactions between transformations and outliers.

of missing values in relation to the data. The missing data mechanisms are missing completely at random (MCAR), missing at random (MAR) and missing not at random data (MNAR). The effects of missing responses in the data set on the usage of transformations have not yet been extensively studied. However, if it is reasonable that the missing data mechanism is MAR, the missing values can be ignored and maximum likelihood theory can be used in combination with a transformation (Rubin, 1976; Lipsitz et al., 2000).

How can the transformations be used when the data is multimodal?

The transformations presented in this paper most likely do not ensure the correction of assumptions when data is multimodal. For instance, a specific variable (e.g., gender or income) can generate different groups in the data with distinct distributions (Bradley, 1977). Therefore, before an appropriate transformation is selected, this effect should be removed or corrected by including a factor as an explanatory variable in the regression model. After this, the residuals should be unimodal. This conventional technique in general modelling theory is also a type of transformation (Fink, 2009).

How does the range of the variable limit the choice of transformations?

One of the most important features that we have to know when choosing a transformation is the range of the variable. Most of the transformations are not mathematically defined for zero or negative values. In order to deal with this problem, three general solutions regarding the use of transformations have been published on this topic. Firstly, the researcher can shift the data with a fixed constant (usually equal to one) or a fixed parameter that makes the data positive. However, using an arbitrary parameter for making the data positive affects the analysis results (Fletcher et al., 2005). Osborne (2002) suggests that adding a constant to the outcome variable only changes the mean and not the other moments of the distribution, and he recommends its use. Atkinson (1987) dedicates a whole chapter to discussing the implications of using this family of transformations with a shifted parameter on model fitting, in particular in the constant parameter and the estimation transformation parameter. Additionally, Hill (1963) and Yeo and Johnson (2000) suggest that asymptotic results of maximum likelihood theory may not hold including a shift parameter. Therefore, the second solution is to use a transformation that includes in its functional form the possibility of using negative and non-negative responses. Finally, Burbidge and Robb (1985) propose to shrink any zero values toward forward zero, while holding the rest constant and applying the maximum likelihood theory.

How are the effects of many zeros in the variable on the transformation?

Data containing a substantial proportion of zeros is commonly known as a zero inflation problem or as an excess zeros problem. If this phenomenon is not correctly handled, the relation between the conditional variance and the dependent variable is not equal, but greater. This problem is called overdispersion and this can lead to an underestimation of the standard errors. Furthermore, when the zero inflated problem is present, transformations may not be applicable to achieve linearity. Another typical situation occurs when changing negative values in the out-

come variable to numbers close to zero. Magee (1988) studied the effects of this change in the outcome variable for the Box-Cox transformation. In this case, the Jacobian of the transformation (see estimation methods in Section 5.2) usually tends to plus or minus infinity (MacKinnon and Magee, 1990) and the transformation parameter tends to be also zero. Furthermore, if a Box-Cox transformation is applied under this condition, the transformed variable will be bounded from below, which is not optimal, if the aim is to deal with non Gaussian assumptions.

How can we decide which variables should be transformed?

Mosteller and Tukey (1977) propose a ladder of transformations to guide the selection of a transformation that helps to fulfill the linearity assumption (see Section 5.2). If it becomes evident that a serious problem in the residuals is present, a transformation in the dependent variable is suggested. Otherwise, if the residuals are well behaved, transforming the outcome can artificially lead to a violation of assumptions, especially to heteroscedasticity. In this case, one or more of the explanatory variables should be transformed (Cohen et al., 2014; Brown, 2015). Box and Tidwell (1962) suggest the use of a power transformation in the explanatory variables in order to linearize the relationship with the outcome variable. The method is known as the Box-Tidwell transformation and seeks to find the optimal transformation parameter under a Box-Cox transformation for each variable that can be transformed (e.g., not for dummy variables). The estimation process is based on maximum likelihood theory and is iterated until convergence. Furthermore, it does not affect the variance stabilization and the Gaussian assumptions of the error term distribution (Box and Cox, 1964).

It is also possible to transform both sides of the regression model. This can be useful when there is a fair certainty that the regression model already describes well the studied interaction, but the assumptions over the error terms are not yet met. In this case, a transformation family T can be applied on both sides of the equation, which leads to the transform both sides (TBS) model:

$$T(y_i) = T[f(x_i, \beta), \lambda] + e_i,$$

and for $f(\cdot)$, the error terms are usually assumed to be additive. The transformation function may take different functional forms. It can simply be the logarithmic transformation, but can also be a more elaborate family of power transformations, such as the Box-Cox. For instance, when the logarithmic transformation is applied on both sides, the level-level regression specification is known as log-log transformation. If done properly, transforming both sides makes the estimation of β more efficient (Carroll and Ruppert, 1988). If the transformation relies on a transformation parameter, adjustments for the estimation of this parameter are suggested. Regarding the Box-Cox transformation, Carroll and Ruppert (1988) propose writing the maximum likelihood function in terms of β , σ_e^2 and λ , and then maximizing it by employing an optimization technique such as the Newton algorithm. As they also acknowledge, it is not always possible to carry out this procedure as it can become computationally expensive. Carroll and Ruppert (1988) suggest two alternatives. One of them is known as the profile likelihood, which is based on the same theory proposed in Box and Cox (1964). The second method is the use of the pseudo-regression model. In terms of parsimony, Carroll and Ruppert (1988) favor the use of the pseudo-regression model method over the profile likelihood. However,

the pseudo-regression model method can have irremediable convergence problems, and when that happens the profile likelihood method is more reliable. Further estimation methods for the TBS method are also studied by Ruppert and Aldershof (1989), Kettl (1991), Nychka and Ruppert (1995) and Wang and Ruppert (1995). In order to calculate standard errors, Carroll and Ruppert (1988) classify six techniques according to the estimation method employed for σ_e^2 , λ and β and which model is fitted to the data.

5.4 Conclusions and future research directions

As this review of transformations shows, the application of transformations is a helpful tool for achieving model assumptions for the linear and linear mixed regression models. In this work, special attention has been paid to the wide range of transformations useful for achieving model assumptions and estimation methods that can be used for the estimation of transformation parameters. We explored the implications of these assumptions, their importance, and the consequences of their violation in terms of estimation and inference. Moreover, an attempt was made to present possible solutions to correct in the case that any of these assumptions is violated. By doing so we showed that transformations can work as a solution for some of these violations; particularly, for non-normality, heteroscedasticity, and non-linearity. In order to combat the misuse of transformations, this work also provides a guide for the correct and thoughtful application.

Because an increasing number of researchers are using the linear and linear mixed regression models, more theory of transformations for these models should be developed in future. For instance, one drawback of transformations is still the interpretation of model results. Interpreting estimations in the transformed scale is not always desired, and most researchers prefer to take decisions on the original scale. Manning (1998, p. 285) summarized this issue by pointing out that “First Bank will not cash a check for log dollars”. Therefore, further research is needed to investigate the bias of back-transforming into the original scale and the interpretation of model results under transformations. Nonetheless, these limitations should be seen as future opportunities. Finally, more effort should be put into the comparison of different estimations under diverse data circumstances.

Chapter 6

The R package **trafo** for transforming linear regression models

6.1 Introduction

To study the relation between two or more variables, the linear regression model is one of the most employed statistical methods. For an appropriate usage of this model, a set of assumptions needs to be fulfilled. These assumptions are, among others, related to the functional form and to the error terms, such as linearity and homoscedasticity. However, in practical applications, these assumptions are not always satisfied. This leads to the question of how the practitioner can move on with the analysis in such case. One way to proceed is to conduct the analysis ignoring the model assumption violations which is, of course, not recommended as it would likely yield misleading results. Another solution is to use more complex methods such as generalized linear regression models or non-parametric methods, as they might fit the data and problem better. A third method, which also constitutes the focus of the present paper, is the application of suitable transformations. Transformations have the potential to correct certain violations and by doing so, enable to continue the analysis with the known (linear) regression model. Due to its convenience, transformations such as the logarithm or the Box-Cox are commonly applied in many branches of sciences; for example in economics (Hossain, 2011) and neuroscience (Morozova et al., 2016). In order to simplify the choice and the usage of transformations in the linear regression model, the R (R Core Team, 2018) package **trafo** (Medina et al., 2018) is developed. The present work is inspired by the framework proposed in Rojas-Perilla (2018, pp. 9-45) and extends other existing R packages that provide transformations.

Many packages that contain transformations do not focus especially on the usage of transformations (Venables and Ripley, 2002; Fox and Weisberg, 2011; Molina and Marhuenda, 2015; Ribeiro Jr. and Diggle, 2016). Therefore, they often only include popular transformations like the logarithmic or the Box-Cox transformation family. The package **car** (Fox and Weisberg, 2011) expands the selection of transformations. It includes the Box-Cox, the basic power, and the Yeo-Johnson transformation families, and uses the maximum likelihood approach for the estimation of the transformation parameter. An exponential transformation proposed by Manly (1976) is provided in the package **caret** (Kuhn, 2008) and the multiple parameter Johnson transformation in the packages **Johnson** (Fernandez, 2014) and **jtrans** (Wang,

2015). While package **MASS** (Venables and Ripley, 2002) and package **car** (Fox and Weisberg, 2011) only provide the maximum likelihood approach for the estimation of the transformation parameter for the Box-Cox family, the estimation can be conducted by a wide range of methods in the **AID** package (Dag et al., 2017). Most of the provided methods are based on goodness of fit tests like the Shapiro-Wilk or the Anderson-Darling test. However, the **AID** package only contains the Box-Cox transformation.

It is noticeable that none of the above-mentioned packages helps the user in the process of deciding which transformation is actually suitable according to the users needs. Furthermore, most packages do not provide tools to see at the first sight if the transformation improves the untransformed model with regards to fulfilling the model assumptions. Therefore, package **trafo** combines and extends the features provided by the packages mentioned above. Additionally to transformations that are already provided by existing packages, the **trafo** package includes, among others, the Bickel-Doksum (Bickel and Doksum, 1981), modulus (John and Draper, 1980), the neglog (Whittaker et al., 2005) and glog (Durbin et al., 2002) transformations that are modifications of the Box-Cox and the logarithmic transformation, respectively, in order to deal with negative values in the response variable. Furthermore, the selection of estimation methods for the transformation parameter is enlarged by methods based on moments and divergence measures (see e.g., Taylor, 1985; Yeo and Johnson, 2000; Royston et al., 2011). The main benefits of the package **trafo** can be summarized as follows:

- An initial check can be conducted that helps to decide if and which transformation is useful for the researchers needs.
- The untransformed model and a model with a transformed dependent variable as well as two transformed models can be run simultaneously, and thus the models can be easily compared with regard to the model assumptions.
- Extensive diagnostics are provided in order to check if the transformation helps to fulfill the model assumptions normality, homoscedasticity, and linearity.

The remainder of this paper is structured as follows. In Section 6.2, the transformations and estimation methods included in the package are presented. Section 6.3 demonstrates in form of a case study the functionality of the package. Section 6.4 summarizes the user-defined function feature of the package. In Section 6.5, some concluding remarks and potential extensions of the package are discussed. Finally, Appendix D.1 presents the mathematical derivations underlying the package.

6.2 Transformations and estimation methods

The equation describing and summarizing the relationship between a continuous outcome variable y and different covariates x (either discrete or continuous) is defined by $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i$, with $i = 1, \dots, n$. This is also known as the linear regression model and is composed by a deterministic and a random component, which rely on different assumptions. Among others, these assumptions can be summarized as follows:

- Normality (N): The conditional distribution of y given x follows a normal distribution. This is an optional, but often desired assumption.
- Homoscedasticity (H): The conditional variance of y given x is constant.
- Linearity (L): The conditional expectation of the outcome variable y given the covariates x is a linear function in x .

As already mentioned, different approaches have been proposed for achieving these model assumptions. Some of them include using alternative estimation methods of the regression terms or applying more complex regression models (see e.g., Nelder and Wedderburn, 1972; Berry, 1993). In this paper, we focus on defining a parsimonious re-specification for the model, such as the usage of non-linear transformations of the outcome variable. The transformations implemented in the package **trafo** basically help to achieve normality. However, most of them simultaneously correct other assumptions (see also Table 6.1 and Table 6.2).

The transformations can be classified into transformations without a transformation parameter and data-driven transformations with a transformation parameter that needs to be estimated. The first set of transformations presented in Table 6.1 comprises, among others, the logarithmic transformation, which is considered due to its popularity and straightforward application. The

Table 6.1: Transformations without transformation parameter.

Transformation	Source	Formula	Support	N	H	L
Log (shift)	Box and Cox (1964)	$\log(y + s)$	$y \in \mathbb{R}$	✗	✗	✗
Glog	Durbin et al. (2002)	$\log(y + \sqrt{y^2 + 1})$	$y \in \mathbb{R}$	✗	✗	✗
Neglog	Whittaker et al. (2005)	$\text{Sign}(y) \log(y + 1)$	$y \in \mathbb{R}$	✗	✗	
Reciprocal	Tukey (1977)	$\frac{1}{y}$	$y \neq 0$	✗	✗	

data-driven transformations presented in Table 6.2 are dominated by the Box-Cox transformation and its modifications or alternatives, e.g., the modulus or Bickel-Doksum transformation. However, more flexible versions of the logarithmic transformation, as the log-shift opt, or the Manly transformation, which is an exponential transformation, are also included in the package **trafo**.

Table 6.1 and 6.2 provide information about the range of the dependent variable that is supported by the transformation. Some transformations are only suitable for positive values of y . This is generally true for the logarithmic and Box-Cox transformations. However, in case that the dependent variable contains negative values, the values are shifted by a deterministic shift s such that $y + s > 0$ by default in package **trafo**. Furthermore, the tables emphasize which assumptions the transformation helps to achieve. These are general suggestions and the actual success always also depends on the data. For specific properties of each transformation we refer to the original references. The square root shift transformation with a data-driven shift in analogy to the log-shift opt transformation is, to the best of our knowledge, firstly implemented in this work. In contrast, a square root transformation with deterministic shift, for example, is suggested in Bartlett (1947).

Since the transformations in Table 6.2 contain transformation parameters that need to be estimated, package **trafo** contains different methodologies for this estimation. The benefit of

Table 6.2: Data-driven transformations.

Transformation	Source	Formula	Support	N	H	L
Box-Cox (shift)	Box and Cox (1964)	$\begin{cases} \frac{(y+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y+s) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	\times	\times	\times
Log-shift opt	Feng et al. (2016)	$\log(y + \lambda)$	$y \in \mathbb{R}$	\times	\times	\times
Bickel-Doksum	Bickel and Doksum (1981)	$\frac{ y ^\lambda \text{Sign}(y) - 1}{\lambda}$ for $\lambda > 0$	$y \in \mathbb{R}$	\times	\times	
Yeo-Johnson	Yeo and Johnson (2000)	$\begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0; \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0; \\ \frac{(1-y)^{2-\lambda} - 1}{\lambda - 2} & \text{if } \lambda \neq 2, y < 0; \\ -\log(1-y) & \text{if } \lambda = 2, y < 0. \end{cases}$	$y \in \mathbb{R}$	\times	\times	
Square Root (shift)	Medina et al. (2018)	$\sqrt{y + \lambda}$	$y \in \mathbb{R}$	\times	\times	
Manly	Manly (1976)	$\begin{cases} \frac{e^{\lambda y} - 1}{\lambda} & \text{if } \lambda \neq 0; \\ y & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	\times	\times	
Modulus	John and Draper (1980)	$\begin{cases} \text{Sign}(y) \frac{(y +1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \text{Sign}(y) \log(y + 1) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	\times		
Dual	Yang (2006)	$\begin{cases} \frac{(y^\lambda - y^{-\lambda})}{2\lambda} & \text{if } \lambda > 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	\times		
Gpower	Kelmansky et al. (2013)	$\begin{cases} \frac{(y + \sqrt{y^2 + 1})^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y + \sqrt{y^2 + 1}) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	\times		

each estimation method depends on the research analysis and the underlying data. They can be summarized as follows:

- Maximum likelihood theory
- Distribution moments optimization: Skewness or kurtosis
- Divergence minimization: Following Kolmogorov-Smirnov (KS), Cramér-von-Mises (KM) or Kullback-Leibler (KL) measurements

The maximum likelihood estimation method finds the set of values for the transformation parameter that maximizes the likelihood function of the dataset under the selected transformation (Box and Cox, 1964). This is a standard approach that is also implemented in several of the mentioned R packages (Venables and Ripley, 2002; Fox and Weisberg, 2011). However, since the maximum likelihood estimation is rather sensitive to outliers, the skewness or kurtosis optimization might be preferable for the estimation of the transformation parameter in the presence of such outliers (see e.g., Royston et al., 2011). These methods are especially favorable when it is important in the analysis to meet these moments. For instance, skewness minimization should be used when it is important to get a symmetric distribution. Additionally, if the focus lies on comparing the whole distribution of the transformed data with a normal distribution, and not only some moments, different divergence measures as the KS, KM or KL can be used (see e.g., Yeo and Johnson, 2000). For all estimation methods, a lambda range on which the functions are evaluated needs to be proposed. Therefore, default values are set for the predefined transformations. For more information about different estimation methods we refer to Rojas-Perilla (2018, pp. 9-45).

Since the user can only decide if the transformation is helpful by checking the above mentioned assumptions, the package **trafo** contains a wide range of diagnostic checks (e.g., Shapiro

Table 6.3: Diagnostic checks provided in the package **trafo**.

Assumption	Diagnostic check	Fast check
Normality	Skewness and kurtosis	X
	Shapiro-Wilk test	X
	Quantile-quantile plot	
	Histograms	
Homoscedasticity	Breusch-Pagan test	X
	Residuals vs. fitted plot	
	Scale-location	
Linearity	Scatter plots between y and x	X
	Observed vs. fitted plot	

and Wilk, 1965; Breusch and Pagan, 1979). A smaller selection is used in the fast check that helps to decide if a transformation might be useful. Table 6.3 summarizes the implemented diagnostic checks that are simultaneously returned for the untransformed and a transformed model or two differently transformed models and indicates which diagnostics are conducted in the fast check. Additionally, plots are provided that help to detect outliers such as the Cook’s distance plot and influential observations by the residuals vs leverage plot.

Another feature of the package **trafo** is the possibility of defining a customized transformation. Thus, a user can also use the infrastructure of the package for a transformation that suits the individuals needs better than the predefined transformations. However, in this version of the package **trafo**, the user needs to define the transformation and the standardized transformation in order to use this feature. For the derivation of the standardized transformation of all predefined transformations, see Appendix D.1.

6.3 Case study

In order to show the functionality of the package **trafo**, we present – in form of a case study – the steps a user faces when checking the assumptions of the linear model. For this illustration, we use the data set called `University` from the R package **Ecdat** (Croissant, 2016). This data set contains variables about the equipment and costs of university teaching and research and can be obtained as follows:

```
R> library(Ecdat)
R> data(University)
```

A practical question for the head of a university could be how study fees (`stfees`) raise the universities net assets (`nassets`). Both variables are metric. Thus, a linear regression could help to explain the relation between these two variables. A linear regression model can be conducted in R using the `lm` function.

```
R> linMod <- lm(nassets ~ stfees, data = University)
```

The features in the package **trafo** that help to find a suitable transformation for this model and to compare different models are summarized in Table 6.4 and illustrated in the next sections.

Table 6.4: Core functions of package **trafo**.

Function	Description
<code>assumptions()</code>	Enables a fast check which transformation is suitable.
<code>trafo_lm()</code>	Compares the untransformed model with a transformed model.
<code>trafo_compare()</code>	Compares two differently transformed models.
<code>diagnostics()</code>	Returns information about the transformation and different diagnostics checks in form of tests.
<code>plot()</code>	Returns graphical diagnostics checks.

Table 6.5: Arguments of function `assumptions`.

Argument	Description	Default
<code>object</code>	Object of class <code>lm</code> .	
<code>method</code>	Estimation method for the transformation parameter.	Maximum likelihood
<code>std</code>	Normal or scaled transformation.	Normal
<code>...</code>	Additional arguments can be added, especially for changing the lambda range for the estimation of the parameter, e.g., <code>manly_lr = c(0.000005, 0.00005)</code>	Default values of lambda range of each transformation

6.3.1 Finding a suitable transformation

It is well known that the reliability of the linear regression model depends on assumptions. Amongst others, normality, homoscedasticity, and linearity are assumed. In this section, we focus on presenting how the user can decide and assess, if and which, transformations help to fulfill these model assumptions. Thus, a first fast check of these model assumptions can be used in the package **trafo** in order to find out if the untransformed model meets these assumptions or if using a transformation seems suitable. The fast check can be conducted by the function `assumptions`. This function returns the skewness, the kurtosis and the Shapiro-Wilk test for normality, the Breusch-Pagan test for homoscedasticity and scatter plots between the dependent and the explanatory variables for checking the linear relation. All possible arguments of the function `assumptions` are summarized in Table 6.5. In the following, we only show the returned normality and homoscedasticity tests. The results are ordered by the highest p value of the Shapiro-Wilk and Breusch-Pagan test.

```
R> assumptions(linMod)
```

The default `lambdarange` for the log shift `opt` transformation is calculated dependent on the data range. The lower value is set to `-2035.751` and the upper value to `404527.249`

The default `lambdarange` for the square root shift transformation is calculated dependent on the data range. The lower value is set to `-2035.751` and the upper value to `404527.249`

Test normality assumption

	Skewness	Kurtosis	Shapiro_W	Shapiro_p
logshiftopt	-0.4201	4.0576	0.9741	0.2132
boxcox	-0.4892	4.2171	0.9621	0.0527
bickeldoksum	-0.4892	4.2171	0.9621	0.0527
gpower	-0.4892	4.2171	0.9621	0.0527
modulus	-0.4892	4.2171	0.9621	0.0527
yeojohnson	-0.4892	4.2171	0.9621	0.0527
dual	-0.4837	4.2180	0.9619	0.0519
sqrtshift	0.6454	5.2752	0.9504	0.0139
log	-1.1653	5.1156	0.9140	0.0004
neglog	-1.1651	5.1150	0.9140	0.0004
glog	-1.1653	5.1156	0.9140	0.0004
untransformed	2.4503	12.7087	0.7922	0.0000
reciprocal	-3.7260	19.0487	0.5676	0.0000

Test homoscedasticity assumption

	BreuschPagan_V	BreuschPagan_p
modulus	0.1035	0.7477
yeojohnson	0.1035	0.7477
boxcox	0.1035	0.7476
bickeldoksum	0.1036	0.7476
gpower	0.1035	0.7476
dual	0.1128	0.7369
logshiftopt	0.1154	0.7341
neglog	0.7155	0.3976
log	0.7158	0.3975
glog	0.7158	0.3975
reciprocal	1.6109	0.2044
sqrtshift	5.4624	0.0194
untransformed	9.8244	0.0017

Following the Shapiro-Wilk test, the best transformation to fulfill the normality assumption is the log-shift opt transformation followed by the Box-Cox, Bickel-Doksum, gpower, modulus and Yeo-Johnson transformation. The similarity or even equality of the test results for different transformations is due to the same functional form in the case of a positive λ and positive values as e.g., the Box-Cox and Bickel-Doksum transformation, or to the rounding at four decimals. For improving the homoscedasticity assumption, all transformations help except the square root (shift) transformation. As mentioned before, default values for the lambda range for all transformations are predefined and these are used in this fast check. Since the default values for the log-shift opt and square root (shift) transformation depend on the range of the response variable, the chosen range is reported in the return. The Manly transformation is not in the list since the default lambda range for the estimation of the transformation parameter

Table 6.6: Arguments of function `trafo_lm`.

Argument	Description	Default
<code>object</code>	Object of class <code>lm</code> .	
<code>trafo</code>	Selected transformation.	Box-Cox
<code>lambda</code>	Estimation or a self-selected numeric value.	Estimation
<code>method</code>	Estimation method for the transformation parameter.	Maximum likelihood
<code>lambdarange</code>	Determines <code>lambdarange</code> for the estimation of the transformation parameter.	Default <code>lambdarange</code> for each transformation.
<code>std</code>	Normal or scaled transformation.	Normal
<code>custom_trafo</code>	Add customized transformation.	None

is not suitable for this data set. It does not fit since the Manly transformation is an exponential transformation and therefore it rather fits for flat or left-skewed data in contrast to most of the other transformations. In the case that the default lambda range does not work, the user can change the lambda range for the transformations manually. Similarly, the user can change the estimation methods for the transformation parameter. For instance, if symmetry is of special interest for the user the skewness minimization might be a better choice than the default maximum likelihood method. In this case study, all assumptions are assumed to be equally important. Thus, we choose the Box-Cox transformation for the further illustrations even though some other transformations would be suitable as well.

6.3.2 Comparing the untransformed model with a transformed model

For a more detailed comparison of the transformed model with the untransformed model, a function called `trafo_lm` (for the arguments see Table 6.6) can be used as follows:

```
R> linMod_trafo <- trafo_lm(linMod)
```

The Box-Cox transformation is the default option such that only the `lm` object needs to be given to the function. The object `linMod_trafo` is of class `trafo_lm` and the user can conduct the methods `print`, `summary` and `plot` in the same way as for an object of class `lm`. The difference is that the new methods simultaneously return the results for both models, the untransformed model and the transformed model. Furthermore, a method called `diagnostics` helps to compare results of normality and homoscedasticity tests. In the following, we will show the return of the `diagnostics` method and some selected plots in order to check the normality, homoscedasticity and the linearity assumption of the linear model.

```
R> diagnostics(linMod_trafo)
```

```
Diagnostics: Untransformed vs transformed model
```

```
Transformation:  boxcox
Estimation method:  ml
Optimal Parameter:  0.1894257
```


Residual diagnostics:

Normality:

Pearson residuals:

	Skewness	Kurtosis	Shapiro_W	Shapiro_p
Untransformed model	2.4503325	12.708681	0.7921672	6.024297e-08
Transformed model	-0.4892222	4.217105	0.9620688	5.267566e-02

Heteroscedasticity:

	BreuschPagan_V	BreuschPagan_p
Untransformed model	9.8243555	0.00172216
Transformed model	0.1035373	0.74762531

The first part of the return shows information of the applied transformation. As chosen, the Box-Cox transformation is used with the optimal transformation parameter around 0.19 which is estimated using the maximum likelihood approach that is also set as default. The optimal transformation parameter differs from 0, which would be equal to the logarithmic transformation, and 1, which means that no transformation is optimal. The Shapiro-Wilk test rejects normality of the residuals of the untransformed model but it does not reject normality for the residuals of the transformed model on a 5% level of significance. Furthermore, the skewness shows that the residuals in the transformed model are more symmetric and the kurtosis is closer to 3, the value of the kurtosis of the normal distribution. The results of the Breusch-Pagan test clearly show that homoscedasticity is rejected in the untransformed model but not in the transformed model. These two findings can be supported by diagnostic plots shown in Figure 6.1.

```
R> plot(linMod_trafo)
```

In order to evaluate the linearity assumption, scatter plots of the dependent variable against the explanatory variable can help. Figure 6.2 shows that the assumption of linearity is violated in the untransformed model. In contrast, the relation between the transformed net assets and the study fees seems to be linear.

As demonstrated above, the user can receive diagnostics for an untransformed and a transformed model with only a little more effort in comparison to fitting the standard linear regression model without transformation. While we only show the example with the default transformation, the user can also easily change the transformation and the estimation method. For instance, the user could choose the log-shift opt transformation with the skewness minimization as estimation method.

```
R> linMod_trafo2 <- trafo_lm(object = linMod,
+   trafo = "logshiftopt", method = "skew")
```

6.3.3 Comparing two transformed models

The user can also compare different transformations with regard to meet the model assumptions. In many present-day applications, the logarithm is often used without longer considera-

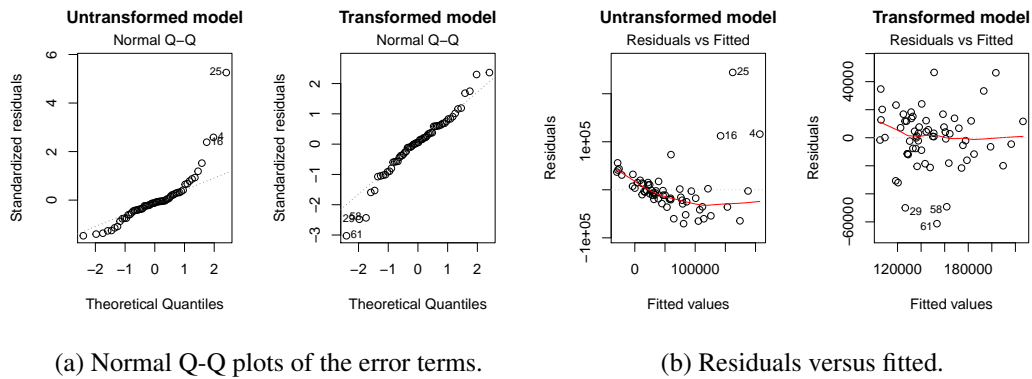


Figure 6.1: Selection of diagnostic plots obtained by using `plot(linMod_trafo)`. (a) shows normal Q-Q plots error terms of the untransformed and the transformed model. (b) shows the residuals against the fitted values of the untransformed and the transformed model.

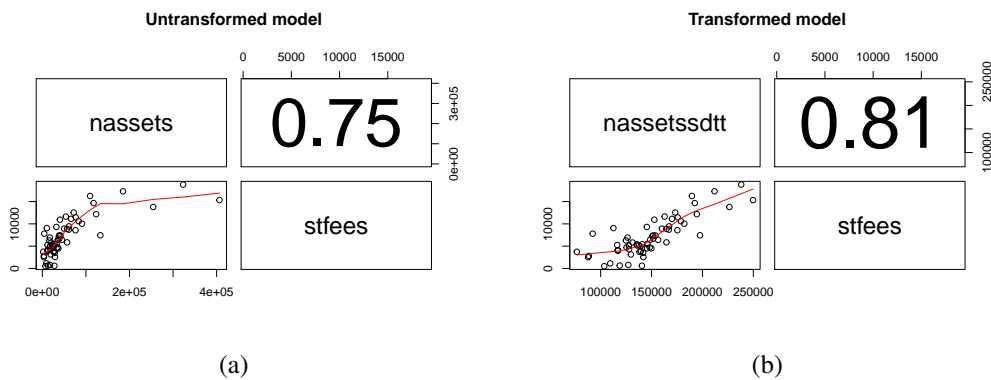


Figure 6.2: Selection of obtained diagnostic plots by using `plot(linMod_trafo)`. (a) shows the scatter plot of the untransformed net assets and the study fees (b) shows scatter plot of the transformed net assets and the study fees. The numbers specify the correlation coefficient between the dependent and independent variable.

Table 6.7: Arguments of function `trafo_compare`.

Argument	Description	Default
<code>object</code>	Object of class <code>lm</code> .	
<code>trafos</code>	List of objects of class <code>trafo</code> .	
<code>std</code>	Normal or scaled transformation.	Normal

tions about its usefulness. In order to compare the logarithm, e.g., with the selected Box-Cox transformation, the user needs to specify two objects of class `trafo` as follows:

```
R> boxcox_uni <- boxcox(linMod)
R> log_uni <- logtrafo(linMod)
```

The utility of `trafo` objects is twofold. First, the user can use the functions for each transformation in order to simply receive the transformed vector. The `print` method gives first information about the vector and the method `as.data.frame` returns the whole data frame with the transformed variable in the last column. The variable is named as the dependent variable with an added `t`.

```
R> head(as.data.frame(boxcox_uni))
```

```
      nassets stfees nassetst
1   3669.71   2821 19.71248
2  12156.00   4037 26.07723
3 185203.00  17296 47.24867
4 323100.00  18800 53.08840
5  32154.00   9314 32.42140
6  41669.00   7388 34.31882
```

Second, the objects can be used to compare linear models with differently transformed dependent variable using function `trafo_compare`. The arguments of this functions are shown in Table 6.7. The user creates an object of class `trafo_compare` by:

```
R> linMod_comp <- trafo_compare(object = linMod,
+   trafos = list(boxcox_uni, log_uni))
```

For this object, the user can use the same methods as for an object of class `trafo_lm`. In this work, we only want to show the return of method `diagnostics`.

```
R> diagnostics(linMod_comp)
```

Diagnostics of two transformed models

```
Transformations:  Box-Cox and Log
Estimation methods:  ml  and no estimation
Optimal Parameters:  0.1894257  and no parameter
```

Residual diagnostics:

Normality:

Pearson residuals:

```
      Skewness Kurtosis Shapiro_W  Shapiro_p
Box-Cox -0.4892222 4.217105 0.9620688 0.0526756632
Log      -1.1653028 5.115615 0.9140135 0.0003534879
```

Heteroscedasticity:

	BreuschPagan_V	BreuschPagan_p
Box-Cox	0.1035373	0.7476253
Log	0.7158162	0.3975197

The first part of the return points out that the Box-Cox transformation is a data-driven transformation with a transformation parameter, while the logarithmic transformation does not adapt to the data. Furthermore, we can see that normality is rejected for the model with a logarithmic transformed dependent variable, while it is not rejected when the Box-Cox transformation is used. The violation of the homoscedasticity assumption can be fixed by both transformations.

6.4 Customized transformation

An additional user-friendly feature in the package **trafo** is the possibility of using the framework also for self-defined transformations. In the following, we show this option for the `glog` transformation.

In a first step, the transformation and the standardized or scaled transformation need to be defined. The mathematical expression of these two functions is presented in the Appendix D.1.2.

```
R> glog_trafo <- function(y) {
+   yt <- log(y + sqrt(y^2 + 1))
+   return(y = yt)}

R> glog_std <- function(y) {
+   zt <- log(y + sqrt(y^2 + 1)) * sqrt(geometric.mean(1 +
+   y^2))
+   return(zt = zt)}
```

Second, the user inserts the two functions as a list argument to the `trafo_lm` function. Furthermore, the user needs to specify for the `trafo` argument if the transformation is without a parameter (`"custom_wo"`) or with one parameter (`"custom_one"`). The `glog` transformation does not rely on a transformation parameter.

```
R> linMod_custom <- trafo_lm(linMod, trafo = "custom_wo",
+   custom_trafo = list(glog_trafo = glog_trafo,
+   glog_std = glog_std))
```

One limitation of this feature is the necessity to insert both the transformation and the scaled transformation since the latter is often not known by the user. Furthermore, the framework is only suitable for transformations without and with one transformation parameter.

6.5 Conclusions and future developments

Even though the development in computing enables the use of complex methods nowadays, transformations are still a parsimonious way to meet model assumptions in a linear regression model. In Section 6.3, we demonstrated how the package **trafo** helps the user to decide easily if and which transformation is suitable to fulfill the model assumptions normality, homoscedasticity and linearity. To the best of our knowledge **trafo** is the only R package that supports this decision process. Furthermore, the package **trafo** provides an extensive collection of transformations usable in linear regression models and a wide range of estimation methods for the transformation parameter. In future versions, we plan to enlarge this collection constantly, also for other types of data, e.g, count data. Additionally, more methods that are available for the class `lm` could be developed for objects of class `trafo_lm`. We would also like to expand the infrastructure for linear mixed regression models.

Appendix D

D.1 Likelihood derivation of the transformations

D.1.1 Log (shift) transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^* . In order to obtain z_i^* , the scaled log (shift) transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{LS} . Therefore, the Jacobian, the scaled, and the inverse of the log (shift) transformation are given below.

The log (shift) transformation presented in Table 6.1 is defined as:

$$y_i^* = \log(y_i + s).$$

In case, the fixed shift parameter s would not be necessary, the standard logarithm function (logarithmic transformation with $s = 0$) is applied.

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{LS} = \left[\prod_{i=1}^n y_i + s \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^n \frac{1}{y_i + s} \\ &= \bar{y}_{LS}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = \log(y_i + s) \bar{y}_{LS}.$$

The inverse function of the log (shift) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \log(y_i + s) \\ y_i^* &= \log(y_i + s) \\ y_i &= e^{y_i^*} - s \\ \Rightarrow f^{-1}(y_i^*) &= e^{y_i^*} - s. \end{aligned}$$

D.1.2 Glog transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^* . In order to obtain z_i^* , the scaled glog transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{GL} . Therefore, the Jacobian, the scaled, and the inverse of the glog transformation are given below.

The glog transformation presented in Table 6.1 is defined as:

$$y_i^* = \log\left(y_i + \sqrt{y_i^2 + 1}\right) \text{ if } \lambda = 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{GL} = \left[\prod_{i=1}^n 1 + y_i^2 \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^n \frac{1}{y_i + \sqrt{y_i^2 + 1}} \left(1 + \frac{2y_i}{2\sqrt{y_i^2 + 1}} \right) \\ &= \prod_{i=1}^n \frac{1}{y_i + \sqrt{y_i^2 + 1}} \left(\frac{y_i + \sqrt{y_i^2 + 1}}{\sqrt{y_i^2 + 1}} \right) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{y_i^2 + 1}} \\ &= \bar{y}_{GL}^{-\frac{n}{2}}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = \log\left(y_i + \sqrt{y_i^2 + 1}\right) \bar{y}_{GL}^{\frac{1}{2}}.$$

The inverse function of the glog transformation is denoted as:

$$\begin{aligned}
f(y_i) &= \log \left(y_i + \sqrt{y_i^2 + 1} \right) \\
y_i^* &= \log \left(y_i + \sqrt{y_i^2 + 1} \right) \\
e^{y_i^*} - y_i &= \sqrt{y_i^2 + 1} \\
(e^{y_i^*} - y_i)^2 &= y_i^2 + 1 \\
e^{y_i^{*2}} - 2e^{y_i^*} y_i &= 1 \\
y_i &= -\frac{(1 - e^{y_i^{*2}})}{2e^{y_i^*}} \\
\Rightarrow f^{-1}(y_i^*) &= -\frac{(1 - e^{y_i^{*2}})}{2e^{y_i^*}}.
\end{aligned}$$

D.1.3 Neglog transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^* . In order to obtain z_i^* , the scaled neglog transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{NL} . Therefore, the Jacobian, the scaled, and the inverse of the neglog transformation are given below.

The neglog transformation presented in Table 6.1 is defined as:

$$y_i^* = \text{sign}(y_i) \log(|y_i| + 1).$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{NL} = \left[\prod_{i=1}^n (|y_i| + 1) \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned}
J(\mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*}{dy} \\
&= \prod_{i=1}^n \text{sign}(y_i) \frac{1}{|y_i| + 1} \\
&= \text{sign} \left(\prod_{i=1}^n y_i \right) \left(\prod_{i=1}^n |y_i| + 1 \right)^{-1} \\
&= \text{sign} \left(\prod_{i=1}^n y_i \right) \bar{y}_{NL}^{-n}.
\end{aligned}$$

The scaled transformation is given by:

$$z_i^* = \text{sign}(y_i) \log(|y_i| + 1) \text{sign} \left(\prod_{i=1}^n y_i \right) \bar{y}_{NL}.$$

The inverse function of the neglog transformation is denoted as:

$$\begin{aligned} f(y_i) &= \text{sign}(y_i) \log(|y_i| + 1) \\ y_i^* &= \text{sign}(y_i) \log(|y_i| + 1) \\ |y_i| &= e^{\text{sign}(y_i^*)y_i^*} - 1 \\ \Rightarrow f^{-1}(y_i^*) &= \pm [e^{\text{sign}(y_i^*)y_i^*} - 1]. \end{aligned}$$

D.1.4 Reciprocal transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^* . In order to obtain z_i^* , the scaled reciprocal transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_R . Therefore, the Jacobian, the scaled, and the inverse of the reciprocal transformation are given below.

The reciprocal transformation presented in Table 6.1 is defined as:

$$y_i^* = \frac{1}{y_i}.$$

The definition of the geometric mean is:

$$\bar{y}_R = \left[\prod_{i=1}^n y_i \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*}{dy_i} \\ &= \prod_{i=1}^n -\frac{1}{y_i^2} \\ &= -\bar{y}_R^{-2n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = -\frac{1}{y_i} \bar{y}_R^2.$$

The inverse function of the reciprocal transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{1}{y_i} \\ y_i^* &= \frac{1}{y_i} \\ y_i &= \frac{1}{y_i^*} \\ \Rightarrow f^{-1}(y_i^*) &= \frac{1}{y_i^*}. \end{aligned}$$

D.1.5 Box-Cox (shift) transformation

$$y_i^*(\lambda) = \begin{cases} \frac{(y_i+s)^{\lambda-1}}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \log(y_i + s) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Box-Cox (shift) transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Box-Cox (shift)(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{BC} . Therefore, the Jacobian, the scaled, and the inverse of the Box-Cox (shift)(A) transformation are given below.

The Box-Cox (shift)(A) transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \frac{(y_i + s)^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0.$$

In case, the fixed shift parameter s is not necessary for making the dataset positive, the standard Box-Cox transformation (with $s = 0$) is applied.

The definition of the geometric mean is:

$$\bar{y}_{BC} = \left[\prod_{i=1}^n y_i + s \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^n \frac{\lambda(y_i + s)^{\lambda-1}}{\lambda} \\ &= \prod_{i=1}^n (y_i + s)^{\lambda-1} \\ &= \bar{y}_{BC}^{n(\lambda-1)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{(y_i + s)^\lambda - 1}{\lambda} \frac{1}{\bar{y}_{BC}^{\lambda-1}}.$$

The inverse function of the Box-Cox (shift)(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ y_i^* &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ y_i &= (\lambda y_i^* + 1)^{\frac{1}{\lambda}} - s \\ \Rightarrow f^{-1}(y_i^*) &= (\lambda y_i^* + 1)^{\frac{1}{\lambda}} - s. \end{aligned}$$

Box-Cox (shift) transformation case (B)

This case is exactly equal to the log (shift) case.

D.1.6 Log-shift opt transformation

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled log-shift opt transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{LSO} . Therefore, the Jacobian, the scaled, and the inverse of the log-shift opt transformation are given below.

The log-shift opt transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \log(y_i + \lambda).$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{LSO} = \left[\prod_{i=1}^n y_i + \lambda \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^n \frac{1}{y_i + \lambda} \\ &= \bar{y}_{LSO}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \log(y_i + \lambda) \bar{y}_{LSO}.$$

The inverse function of the log-shift opt transformation is denoted as:

$$\begin{aligned} f(y_i) &= \log(y_i + \lambda) \\ y_i^* &= \log(y_i + \lambda) \\ y_i &= e^{y_i^*} - \lambda \\ \Rightarrow f^{-1}(y_i^*) &= e^{y_i^*} - \lambda. \end{aligned}$$

D.1.7 Bickel-Docksum transformation

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Bickel-Docksum transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{BD} . Therefore, the Jacobian, the scaled, and the inverse of the Bickel-Docksum transformation are given below.

The Bickel-Docksum transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \frac{|y_i|^\lambda \text{sign}(y_i) - 1}{\lambda} \text{ if } \lambda > 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{BD} = \left[\prod_{i=1}^n |y_i| \right]^{\frac{1}{n}}.$$

Therefore, the expression of the jacobian comes to:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^n \frac{\text{sign}(y_i) \lambda |y_i|^{\lambda-1}}{\lambda} \\ &= \text{sign} \left(\prod_{i=1}^n y_i \right) \left(\prod_{i=1}^n |y_i| \right)^{\lambda-1} \\ &= \text{sign} \left(\prod_{i=1}^n y_i \right) \bar{y}_{BD}^{n(\lambda-1)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{|y_i|^\lambda \text{sign}(y_i) - 1}{\lambda} \frac{1}{\text{sign} \left(\prod_{i=1}^n y_i \right) \bar{y}_{BD}^{(\lambda-1)}}.$$

The inverse function of the Bickel-Docksum transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{|y_i|^\lambda \text{sign}(y_i) - 1}{\lambda} \\ y_i^* &= \frac{|y_i|^\lambda \text{sign}(y_i) - 1}{\lambda} \\ |y_i| &= [\text{sign}(y_i^*)(y_i^* \lambda + 1)]^{\frac{1}{\lambda}} \\ \Rightarrow f^{-1}(y_i^*) &= \pm [\text{sign}(y_i^*)(y_i^* \lambda + 1)]^{\frac{1}{\lambda}}. \end{aligned}$$

D.1.8 Yeo-Johnson transformation

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_i+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y_i \geq 0 \quad (A); \\ \log(y_i + 1) & \text{if } \lambda = 0, y_i \geq 0 \quad (B); \\ -\frac{(1-y_i)^{2-\lambda} - 1}{2-\lambda} & \text{if } \lambda \neq 2, y_i < 0 \quad (C); \\ -\log(1 - y_i) & \text{if } \lambda = 0, y_i < 0 \quad (D). \end{cases}$$

Yeo-Johnson transformation case (A)

This case is exactly equal to the Box-Cox (shift) case (A), with $s = 1$.

Yeo-Johnson transformation case (B)

This case is exactly equal to the log (shift) case, with $s = 1$.

Yeo-Johnson transformation case (C)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Yeo-Johnson(C) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{YC} . Therefore, the Jacobian, the scaled, and the inverse of the Yeo-Johnson(C) transformation are given below.

The Yeo-Johnson(C) transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = -\frac{(1 - y_i)^{2-\lambda} - 1}{2 - \lambda} \text{ if } \lambda \neq 2 \text{ and } y_i < 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{YC} = \left[\prod_{i=1}^n (1 - y_i) \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned}
 J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\
 &= \prod_{i=1}^n \frac{(2-\lambda)(1-y_i)^{1-\lambda}}{2-\lambda} \\
 &= \prod_{i=1}^n (1-y_i)^{1-\lambda} \\
 &= \bar{y}_{YC}^{n(1-\lambda)}.
 \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = -\frac{(1-y_{ij})^{2-\lambda} - 1}{2-\lambda} \bar{y}_{YC}^{n(1-\lambda)}.$$

The inverse function of the Yeo-Johnson(C) transformation is denoted as:

$$\begin{aligned}
 f(y_i) &= -\frac{(1-y_i)^{2-\lambda} - 1}{2-\lambda} \\
 y_i^* &= -\frac{(1-y_i)^{2-\lambda} - 1}{2-\lambda} \\
 -y_i^*(2-\lambda) &= (1-y_i)^{2-\lambda} - 1 \\
 y_i &= 1 - [-y_i^*(2-\lambda) + 1]^{\frac{1}{2-\lambda}} \\
 \Rightarrow f^{-1}(y_i^*) &= 1 - [-y_i^*(2-\lambda) + 1]^{\frac{1}{2-\lambda}}.
 \end{aligned}$$

Yeo-Johnson transformation case (D)

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^* . In order to obtain z_i^* , the scaled Yeo-Johnson(D) transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{YD} . Therefore, the Jacobian, the scaled, and the inverse of the Yeo-Johnson(D) transformation are given below.

The Yeo-Johnson(D) transformation presented in Table 6.2 is defined as:

$$y_i^* = -\log(1-y_i).$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{YD} = \left[\prod_{i=1}^n (1-y_i) \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^n \frac{1}{1 - y_i} \\ &= \bar{y}_{YD}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = -\log(1 - y_i) \bar{y}_{YD}.$$

The inverse function of the Yeo-Johnson(D) transformation is denoted as:

$$\begin{aligned} f(y_i) &= -\log(1 - y_i) \\ y_i^* &= -\log(1 - y_i) \\ y_i &= -e^{-y_i^*} + 1 \\ \Rightarrow f^{-1}(y_i^*) &= -e^{-y_i^*} + 1. \end{aligned}$$

D.1.9 Square root-shift opt transformation

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain z_i^* , the scaled square root-shift opt transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{SR} . Therefore, the Jacobian, the scaled, and the inverse of the square root-shift opt transformation are given below.

The square root-shift opt transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \sqrt{y_i + \lambda}.$$

The definition of the geometric mean is:

$$\bar{y}_{SR} = \left[\prod_{i=1}^n y_i + \lambda \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^n \frac{1}{2\sqrt{y_i + \lambda}} \\ &= \frac{1}{2} \bar{y}_{SR}^{-\frac{n}{2}}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = -\frac{1}{y_i} \bar{y}_{SR}^2.$$

The inverse function of the square root-shift opt transformation is denoted as:

$$\begin{aligned} f(y_i) &= \sqrt{y_i + \lambda} \\ y_i^* &= \sqrt{y_i + \lambda} \\ y_i &= y_i^{*2} - \lambda \\ \Rightarrow f^{-1}(y_i^*) &= y_i^{*2} - \lambda. \end{aligned}$$

D.1.10 Manly transformation

$$y_i^*(\lambda) = \begin{cases} \frac{e^{\lambda y_i} - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ y_i & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Manly transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Manly(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_M . Therefore, the Jacobian, the scaled, and the inverse of the Manly(A) transformation are given below.

The Manly(A) transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \frac{e^{\lambda y_i} - 1}{\lambda} \text{ if } \lambda \neq 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\begin{aligned} \bar{y}_M &= \left[\prod_{i=1}^n e^{y_i} \right]^{\frac{1}{n}} \\ &= \left[e^{\sum_{i=1}^n y_i} \right]^{\frac{1}{n}} \\ &= e^{\bar{y}}. \end{aligned}$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned}
 J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\
 &= \prod_{i=1}^n \frac{\lambda e^{\lambda y_i}}{\lambda} \\
 &= \left(\prod_{i=1}^n e^{y_i} \right)^\lambda \\
 &= \bar{y}_M^{\lambda n} \\
 &= e^{\lambda n \bar{y}}.
 \end{aligned}$$

The scaled transformation is given by:

$$\begin{aligned}
 z_i^*(\lambda) &= \frac{e^{\lambda y_i} - 1}{\lambda} \frac{1}{\bar{y}_M^\lambda} \\
 &= \frac{e^{\lambda y_i} - 1}{\lambda} \frac{1}{e^{\lambda \bar{y}}}.
 \end{aligned}$$

The inverse function of the Manly(A) transformation is denoted as:

$$\begin{aligned}
 f(y_i) &= \frac{e^{\lambda y_i} - 1}{\lambda} \\
 y_i^* &= \frac{e^{\lambda y_i} - 1}{\lambda} \\
 \lambda y_i^* + 1 &= e^{\lambda y_i} \\
 y_i &= \frac{\log(\lambda y_i^* + 1)}{\lambda} \\
 \Rightarrow f^{-1}(y_i^*) &= \frac{\log(\lambda y_i^* + 1)}{\lambda}.
 \end{aligned}$$

Manly transformation case (B)

The variable remains equal, $y_i^* = y_i$.

D.1.11 Modulus transformation

$$y_i^*(\lambda) = \begin{cases} \text{sign}(y_i) \frac{(|y_i|+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \text{sign}(y_i) \log(|y_i| + 1) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Modulus transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled modulus(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{MA} . Therefore, the Jacobian, the scaled, and the inverse of the modulus(A) transformation are given below.

The modulus(A) transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \text{sign}(y_i) \frac{(|y_i| + 1)^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{MA} = \left[\prod_{i=1}^n |y_i| + 1 \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^n \frac{\text{sign}(y_i) \lambda (|y_i| + 1)^{\lambda-1}}{\lambda} \\ &= \text{sign} \left(\prod_{i=1}^n y_i \right) \left(\prod_{i=1}^n |y_i| + 1 \right)^{\lambda-1} \\ &= \text{sign} \left(\prod_{i=1}^n y_i \right) \bar{y}_{MA}^{n(\lambda-1)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \text{sign}(y_i) \frac{(|y_i| + 1)^\lambda - 1}{\lambda} \frac{1}{\text{sign} \left(\prod_{i=1}^n y_i \right) \bar{y}_{MA}^{(\lambda-1)}}.$$

The inverse function of the modulus(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \text{sign}(y_i) \frac{(|y_i| + 1)^\lambda - 1}{\lambda} \\ y_i^* &= \text{sign}(y_i) \frac{(|y_i| + 1)^\lambda - 1}{\lambda} \\ |y_i| &= \left[\text{sign}(y_i^*) \lambda + 1 \right]^{\frac{1}{\lambda}} - 1 \\ \Rightarrow f^{-1}(y_i^*) &= \pm \left[(\text{sign}(y_i^*) \lambda + 1)^{\frac{1}{\lambda}} - 1 \right]. \end{aligned}$$

Modulus transformation case (B)

This case is exactly equal to the neglog transformation case.

D.1.12 Dual power transformation

$$y_i^*(\lambda) = \begin{cases} \frac{y_i^\lambda - y_i^{-\lambda}}{2\lambda} & \text{if } \lambda > 0 \quad (A); \\ \log(y_i) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Dual power transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled dual power(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{DA} . Therefore, the Jacobian, the scaled, and the inverse of the dual power(A) transformation are given below. The dual power(A) transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \frac{y_i^\lambda - y_i^{-\lambda}}{2\lambda} \text{ if } \lambda > 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{DA} = \left[\prod_{i=1}^n \left(y_i^{\lambda-1} + y_i^{-\lambda-1} \right) \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^n \frac{\lambda y_i^{\lambda-1} + \lambda y_i^{-\lambda-1}}{2\lambda} \\ &= \frac{1}{2} \bar{y}_{DA}^n. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{y_i^\lambda - y_i^{-\lambda}}{2\lambda} \frac{2}{\bar{y}_{DA}}.$$

The inverse function of the dual power(A) transformation is found by solving the quadratic by completing the square as:

$$\begin{aligned}
f(y_i) &= \frac{y_i^\lambda - y_i^{-\lambda}}{2\lambda} \\
y_i^* &= \frac{y_i^\lambda - y_i^{-\lambda}}{2\lambda} \\
2\lambda y_i^* &= y_i^\lambda - y_i^{-\lambda} \\
2\lambda y_i^* &= y_i^\lambda - \frac{1}{y_i^\lambda} \\
2\lambda y_i^* &= \frac{y_i^{2\lambda} - 1}{y_i^\lambda} \\
2\lambda y_i^* y_i^\lambda &= y_i^{2\lambda} - 1 \\
1 + \lambda^2 y_i^{*2} &= y_i^{2\lambda} - 2\lambda y_i^* y_i^\lambda + \lambda^2 y_i^{*2} \\
1 + \lambda^2 y_i^{*2} &= (y_i^\lambda - \lambda y_i^*)^2 \\
\sqrt{1 + \lambda^2 y_i^{*2}} + \lambda y_i^* &= y_i^\lambda \\
y_i &= \left[\sqrt{1 + \lambda^2 y_i^{*2}} + \lambda y_i^* \right]^{\frac{1}{\lambda}} \\
\Rightarrow f^{-1}(y_i^*) &= \left[\sqrt{1 + \lambda^2 y_i^{*2}} + \lambda y_i^* \right]^{\frac{1}{\lambda}}.
\end{aligned}$$

Dual power transformation case (B)

This case is exactly equal to the Box-Cox (shift) transformation, case (B).

D.1.13 Gpower transformation

$$y_i^*(\lambda) = \begin{cases} \frac{\left(y_i + \sqrt{y_i^2 + 1} \right)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \log \left(y_i + \sqrt{y_i^2 + 1} \right) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Gpower transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled gpower(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda, y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by \bar{y}_{GA} . Therefore, the Jacobian, the scaled, and the inverse of the gpower(A) transformation are given below.

The gpower(A) transformation presented in Table 6.2 is defined as:

$$y_i^*(\lambda) = \frac{\left[y_i + \sqrt{y_i^2 + 1} \right]^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{GA} = \left[\prod_{i=1}^n \left(y_i + \sqrt{y_i^2 + 1} \right)^{\lambda-1} \left(1 + \frac{y_i}{\sqrt{y_i^2 + 1}} \right) \right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \mathbf{y}) &= \prod_{i=1}^n \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^n \frac{\lambda \left(y_i + \sqrt{y_i^2 + 1} \right)^{\lambda-1} \left(1 + \frac{2y_i}{2\sqrt{y_i^2+1}} \right)}{\lambda} \\ &= \bar{y}_{GA}^n. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{\left[y_i + \sqrt{y_i^2 + 1} \right]^\lambda - 1}{\lambda} \frac{1}{\bar{y}_{GA}}.$$

The inverse function of the gpower(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{\left[y_i + \sqrt{y_i^2 + 1} \right]^\lambda - 1}{\lambda} \\ y_i^* &= \frac{\left[y_i + \sqrt{y_i^2 + 1} \right]^\lambda - 1}{\lambda} \\ \lambda y_i^* + 1 &= \left[y_i + \sqrt{y_i^2 + 1} \right]^\lambda \\ (\lambda y_i^* + 1)^{\frac{1}{\lambda}} &= y_i + \sqrt{y_i^2 + 1} \\ \left[(\lambda y_i^* + 1)^{\frac{1}{\lambda}} - y_i \right]^2 &= \left[\sqrt{y_i^2 + 1} \right]^2 \\ (\lambda y_i^* + 1)^{\frac{2}{\lambda}} - 2y_i(\lambda y_i^* + 1)^{\frac{1}{\lambda}} + y_i^2 &= y_i^2 + 1 \\ -y_i(\lambda y_i^* + 1)^{\frac{1}{\lambda}} &= \frac{1 - (\lambda y_i^* + 1)^{\frac{2}{\lambda}}}{2} \\ y_i &= - \left[\frac{1 - (\lambda y_i^* + 1)^{\frac{2}{\lambda}}}{2(\lambda y_i^* + 1)^{\frac{1}{\lambda}}} \right] \\ \Rightarrow f^{-1}(y_i^*) &= - \left[\frac{1 - (\lambda y_i^* + 1)^{\frac{2}{\lambda}}}{2(\lambda y_i^* + 1)^{\frac{1}{\lambda}}} \right]. \end{aligned}$$

Gpower transformation case (B)

This case is exactly equal to the glog transformation case.

Bibliography

- Agresti, A., B. Caffo, and P. Ohman-Strickland (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Computational Statistics & Data Analysis 47(3), 639–653.
- Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55(1), 117–128.
- Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The R package **laeken**. Journal of Statistical Software 54(15), 1–25.
- Alfons, A., M. Templ, and P. Filzmoser (2010). An object-oriented framework for statistical simulation: The R package **simFrame**. Journal of Statistical Software 37(3), 1–36.
- Ammermüller, A., A. M. Weber, and P. Westerheide (2005). Die Entwicklung und Verteilung des Vermögens privater Haushalte unter besonderer Berücksichtigung des Produktivvermögens. Abschlussbericht zum Forschungsauftrag des Bundesministeriums für Gesundheit und Soziale Sicherung, Zentrum für Europäische Wirtschaftsforschung GmbH.
- Ampudia, M., H. van Vlokhoven, and D. Żochowski (2016). Financial fragility of euro area households. Journal of Financial Stability 27, 250–262.
- Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. Journal of the American Statistical Association 49(268), 765–769.
- Andreasch, M. and P. Lindner (2016). Micro- and macrodata: A comparison of the Household Finance and Consumption Survey with financial accounts in Austria. Journal of Official Statistics 32(1), 1–28.
- Andrews, D. F. (1971). A note on the selection of data transformations. Biometrika 58(2), 249–254.
- Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. Biometrika 35(3/4), 246–254.
- Anscombe, F. J. and I. Guttman (1960). Rejection of outliers. Technometrics 2(2), 123–147.
- Anscombe, F. J. and J. W. Tukey (1963). The examination and analysis of residuals. Technometrics 5(2), 141–160.

- Armitage, P., G. Berry, and J. N. S. Matthews (2008). Statistical Methods in Medical Research. Hoboken: John Wiley & Sons.
- Arndt, O., H. Dalezios, P. Steden, and G. Färber (2009). Die regionale Inzidenz von Bundesmitteln. In H. Mäding (Ed.), Öffentliche Finanzströme und räumliche Entwicklung, pp. 9–48. Hannover: Verlag der Academy for Spatial Research and Planning (ARL).
- Asar, Ö., O. Ilk, and O. Dag (2017). Estimating Box-Cox power transformation parameter via goodness-of-fit tests. Communications in Statistics - Simulation and Computation 46(1), 91–105.
- Atkinson, A. C. (1973). Testing transformations to normality. Journal of the Royal Statistical Society Series B 35(3), 473–479.
- Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables. Journal of the Royal Statistical Society Series B 44(1), 1–36.
- Atkinson, A. C. (1986). Diagnostic tests for transformations. Technometrics 28(1), 29–37.
- Atkinson, A. C. (1987). Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Oxford: Oxford University Press.
- Atkinson, A. C. and M. Riani (2012). Robust Diagnostic Regression Analysis. New York: Springer-Verlag.
- Babu, G. (1986). A note on bootstrapping the variance of sample quantile. Annals of the Institute of Statistical Mathematics 38(3), 439–443.
- Barnett, V. and T. Lewis (1984). Outliers in Statistical Data. Hoboken: John Wiley & Sons.
- Barry, D. (1993). Testing for additivity of a regression function. The Annals of Statistics 21(1), 235–254.
- Bartlett, M. S. (1935). The effect of non-normality on the t distribution. Mathematical Proceedings of the Cambridge Philosophical Society 31(2), 223–231.
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London Series A 160(901), 268–282.
- Bartlett, M. S. (1947). The use of transformations. Biometrics 3(1), 39–52.
- Bartlett, M. S. and D. Kendall (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. Supplement to the Journal of the Royal Statistical Society 8(1), 128–138.
- Barton, K. (2018). MuMIn: Multi-Model Inference. R package version 1.40.4.
- Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using **lme4**. Journal of Statistical Software 67(1), 1–48.

- Battese, G., R. Harter, and W. Fuller (1988). An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83(401), 28–36.
- Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. Biometrika 32(3/4), 243–262.
- Bell, W. R., W. W. Basel, and J. J. Maples (2016). An overview of the U.S. Census Bureau's small area income and poverty estimates program. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 379–403. Hoboken: John Wiley & Sons.
- Bell, W. R. and E. T. Huang (2006). Using the t -distribution to deal with outliers in small area estimation. Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health, U.S. Census Bureau.
- Belsley, D. A., E. Kuh, and R. E. Welsch (2005). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Hoboken: John Wiley & Sons.
- Berry, W. D. (1993). Understanding Regression Assumptions. Thousand Oaks: SAGE Publications.
- Beste, J., M. M. Grabka, and J. Goebel (2018). Armut in Deutschland. AStA Wirtschafts- und Sozialstatistisches Archiv 12(1), 27–62.
- Bhat, C. R. (1994). Imputing a continuous income variable from grouped and missing income observations. Economics Letters 46(4), 311–319.
- Bickel, P. J. and K. A. Doksum (1981). An analysis of transformations revisited. Journal of the American Statistical Association 76(374), 296–311.
- Bischl, B. and M. Lang (2015). **parallelMap**: Unified Interface to Parallelization Back-Ends. R package version 1.3.
- Bivand, R., T. Keitt, and B. Rowlingson (2018). **rgdal**: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.2-18.
- Bivand, R. and N. Lewin-Koh (2017). **maptools**: Tools for Reading and Handling Spatial Objects. R package version 0.9-2.
- Bivand, R., E. Pebesma, and V. Gómez-Rubio (2013). Applied Spatial Data Analysis with R. New York: Springer-Verlag.
- Bivand, R. and C. Rundel (2017). **rgeos**: Interface to Geometry Engine - Open Source ('GEOS'). R package version 0.3-26.
- Bland, J. and D. Altman (1996). Transformations, means, and confidence intervals. BMJ: British Medical Journal 312(7038), 1079.

- Blaylock, J. R. and D. M. Smallwood (1985). Box-Cox transformations and a heteroscedastic error variance: Import demand equations revisited. International Statistical Review 53(1), 91–97.
- Blom, G. (1958). Statistical Estimates and Transformed Beta-Variables. Hoboken: John Wiley & Sons.
- Blum, U., M. Brachert, H.-U. Brautzsch, K. Brenke, H. Buscher, D. Dietrich, W. Dürig, P. Franz, J. Günther, P. Haug, et al. (2011). Wirtschaftlicher Stand und Perspektiven für Ostdeutschland: Studie im Auftrag des Bundesministeriums des Innern. IWH-Sonderheft, Institut für Wirtschaftsforschung Halle.
- Blum, U., H. S. Buscher, H. Gabrisch, J. Günther, G. Heimpold, C. Lang, U. Ludwig, M. T. W. Rosenfeld, and L. Schneider (2010). Ostdeutschlands Transformation seit 1990 im Spiegel wirtschaftlicher und sozialer Indikatoren. IWH-Sonderheft, Institut für Wirtschaftsforschung Halle.
- Bock, R. (1985). Multivariate Statistical Methods in Behavioral Research. Skokie: Scientific Software International.
- Boonstra, H. (2012). hbsae: Hierarchical Bayesian Small Area Estimation. R package version 1.0.
- Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society Series B 26(2), 211–252.
- Box, G. E. P. and D. R. Cox (1982). An analysis of transformations revisited, rebutted. Journal of the American Statistical Association 77(377), 209–210.
- Box, G. E. P. and P. W. Tidwell (1962). Transformation of the independent variables. Technometrics 4(4), 531–550.
- Boylan, T. A., M. P. Cuddy, and I. G. O’Muircheartaigh (1982). Import demand equations: An application of a generalized Box-Cox methodology. International Statistical Review 50(1), 103–112.
- Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. The American Statistician 31(4), 147–150.
- Breidenbach, J. (2015). JoSAE: Functions for some Unit-Level Small Area Estimators and their Variances. R package version 0.2.3.
- Breusch, T. S. and A. R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica 47(5), 1287–1294.
- Brown, G., R. Chambers, P. Heady, and D. Heasman (2001). Evaluation of small area estimation methods: An application to unemployment estimates from the UK LFS. Symposium 2001 - Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada.

- Brown, J. D. (2015). Linear Models in Matrix Form: A Hands-On Approach for the Behavioral Sciences. Basel: Springer International Publishing.
- Brunnermeier, M. and I. Schnabel (2016). Bubbles and central banks: Historical perspectives. In M. D. Bordo, Ø. Eitrheim, M. Flandreau, and J. F. Qvigstad (Eds.), Central Banks at a Crossroads: What can we Learn from History. Cambridge: Cambridge University Press.
- Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963–1987. Journal of Econometrics 65(1), 109–154.
- Budde, R. and L. Eilers (2014). Sozioökonomische Daten auf Rasterebene: Datenbeschreibung der microm-Rasterdaten. RWI Materialien, Leibniz-Institut für Wirtschaftsforschung.
- bulwiengesa AG (2018). Intelligente Daten für klare Entscheidungen. <https://www.bulwiengesa.de/de/leistungsprogramm/daten>. [accessed: 01.09.2018].
- Bundesamt für Eich- und Vermessungswesen (2017). Verwaltungsgrenzen (VGD) - 1:250.000 Bezirksgrenzen, Daten vom 01.04.2017 von SynerGIS. http://data-synergis.opendata.arcgis.com/datasets/bb4acc011100469185d2e59fa4cae5fc_0. [accessed: 07.02.2018].
- Bundesinstitut für Bau-, Stadt-, und Raumforschung (2017). Indikatoren und Karten zur Raum- und Stadtentwicklung. <http://www.inkar.de/>. Datenlizenz Deutschland - Namensnennung - Version 2.0 [accessed: 12.04.2018].
- Burbidge, J. B., L. Magee, and A. L. Robb (1988). Alternative transformations to handle extreme values of the dependent variable. Journal of the American Statistical Association 83(401), 123–127.
- Burbidge, J. B. and A. L. Robb (1985). Evidence on wealth-age profiles in Canadian cross-section data. Canadian Journal of Economics 18(4), 854–875.
- Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research 33(2), 261–304.
- Cameron, M. A. (1984). Choosing a symmetrizing power transformation. Journal of the American Statistical Association 79(385), 107–108.
- Carroll, C. D. (1998). Why do the rich save so much? NBER Working Paper, National Bureau of Economic Research.
- Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. Journal of the Royal Statistical Society Series B 42(1), 71–78.
- Carroll, R. J. (1982a). Tests for regression parameters in power transformation models. Scandinavian Journal of Statistics 9(4), 217–222.
- Carroll, R. J. (1982b). Two examples of transformations when there are possible outliers. Journal of the Royal Statistical Society Series C 31(2), 149–152.

- Carroll, R. J. and D. Ruppert (1981). On prediction and the power transformation family. Biometrika 68(3), 609–615.
- Carroll, R. J. and D. Ruppert (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79(386), 312–313.
- Carroll, R. J. and D. Ruppert (1985). Transformations in regression: A robust analysis. Technometrics 27(1), 1–12.
- Carroll, R. J. and D. Ruppert (1987). Diagnostics and robust estimation when transforming the regression model and the response. Technometrics 29(3), 287–299.
- Carroll, R. J. and D. Ruppert (1988). Transformation and Weighting in Regression. Boca Raton: CRC Press.
- Casas-Cordero, C., J. Encina, and P. Lahiri (2016). Poverty mapping for the Chilean comunas. In M. Pratesi (Ed.), Analysis of Poverty Data by Small Area Estimation, pp. 379–403. Hoboken: John Wiley & Sons.
- Chambers, J. and T. Hastie (1992). Statistical Models in S. Boca Raton: CRC Press.
- Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey (1983). Graphical Methods for Data Analysis. Boca Raton: CRC Press.
- Chambers, R. L. (1986). Outlier robust finite population estimation. Journal of the American Statistical Association 81(396), 1063–1069.
- Chambers, R. L. and N. Tzavidis (2006). M-quantile models for small area estimation. Biometrika 93(2), 255–268.
- Chandra, H., K. Aditya, and S. Kumar (2018). Small-area estimation under a log-transformed area-level model. Journal of Statistical Theory and Practice 12(3), 497–505.
- Chandra, H., N. Salvati, and R. Chambers (2015). A spatially nonstationary Fay-Herriot model for small area estimation. Journal of the Survey Statistics and Methodology 3(2), 109–135.
- Chatterjee, A. (2011). Asymptotic properties of sample quantiles from a finite population. Annals of the Institute of Statistical Mathematics 63(1), 157–179.
- Chatterjee, S. and A. S. Hadi (2015). Regression Analysis by Example. Hoboken: John Wiley & Sons.
- Chen, S. (2002). Rank estimation of transformation models. Econometrica 70(4), 1683–1697.
- Chen, W. W. and R. S. Deo (2004). Power transformations to induce normality and their applications. Journal of the Royal Statistical Society Series B 66(1), 117–130.
- Cheng, T.-C. (2005). Robust regression diagnostics with data transformations. Computational Statistics & Data Analysis 49(3), 875–891.

- Cheung, K. and S. Lee (2005). Variance estimation for sample quantiles using the m out of n bootstrap. Annals of the Institute of Statistical Mathematics 57(2), 279–290.
- Chung, S. H., W. L. Pearn, and Y. S. Yang (2007). A comparison of two methods for transforming non-normal manufacturing data. The International Journal of Advanced Manufacturing Technology 31(9-10), 957–968.
- Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Human Genetics 11(1), 47–52.
- Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics 3(1), 22–38.
- Cohen, J., P. Cohen, S. G. West, and L. S. Aiken (2014). Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Mahwah: Lawrence Erlbaum Associates.
- Conover, W. J. and R. L. Iman (1981). Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician 35(3), 124–129.
- Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics 19(1), 15–18.
- Cook, R. D. and P. Prescott (1981). On the accuracy of Bonferroni significance levels for detecting outliers in linear models. Technometrics 23(1), 59–63.
- Cook, R. D. and P. Wang (1983). Transformations and influential cases in regression. Technometrics 25(4), 337–343.
- Cook, R. D. and S. Weisberg (1982). Residuals and Influence in Regression. New York: Chapman & Hall.
- Council of the European Union (2001). Report on indicators in the field of poverty and social exclusions. Technical report, European Union.
- Cowell, F. (2011). Measuring Inequality. Oxford: Oxford University Press.
- Cramér, H. (1946). Mathematical Methods of Statistics. Princeton: Princeton University Press.
- Croissant, Y. (2016). Ecdat: Data Sets for Econometrics. R package version 0.3-1.
- Curtiss, J. H. (1943). On transformations used in the analysis of variance. The Annals of Mathematical Statistics 14(2), 107–122.
- Cuthbert, D., F. S. Wood, and J. W. Gorman (1971). Fitting Equations to Data: Computer Analysis of Multifactor Data for Scientists and Engineers. Hoboken: John Wiley & Sons.
- da Costa, A. F. and A. F. Crepaldi (2014). The bias in reversing the Box-Cox transformation in time series forecasting: An empirical study based on neural networks. Neurocomputing 136, 281–288.

- Dag, O., O. Asar, and O. Ilk (2017). AID: Box-Cox Power Transformation. R package version 2.3.
- Datta, G. S., M. Ghosh, R. Steorts, and J. Maples (2011). Bayesian benchmarking with applications to small area estimation. Test 20(3), 574–588.
- Datta, G. S. and P. Lahiri (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Statistica Sinica 10(2), 613–627.
- Datta, G. S., P. Lahiri, and T. Maiti (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. Journal of Statistical Planning and Inference 102(1), 83–97.
- David, H. and H. Nagaraja (2003). Order Statistics. Hoboken: John Wiley & Sons.
- Der Paritätische Gesamtverband (2017). Menschenwürde ist Menschenrecht. Bericht zur Armutsentwicklung in Deutschland 2017, Der Paritätische Gesamtverband.
- Deutsche Bundesbank (2016). Vermögen und Finanzen privater Haushalte in Deutschland: Ergebnisse der Vermögensbefragung 2014. Monatsbericht, Deutsche Bundesbank.
- Deutsche Bundesbank (2018). Monatsbericht August 2018. Monatsbericht, Deutsche Bundesbank.
- Dielmann, T., C. Lowry, and R. Pfaffenberger (1994). A comparison of quantile estimators. Communication in Statistics - Simulation and Computation 23(2), 355–371.
- Doksum, K. A. (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79(386), 316–319.
- Doksum, K. A. and C.-W. Wong (1983). Statistical tests based on transformed data. Journal of the American Statistical Association 78(382), 411–417.
- Dougherty, C. (2011). Introduction to Econometrics. Oxford: Oxford University Press.
- Dragulescu, A. A. (2014). xlsx: Read, Write, Format Excel™2007 and Excel™97/2000/XP/2003 Files. R package version 0.5.7.
- Draper, N. and D. Cox (1969). On distributions and their transformation to normality. Journal of the Royal Statistical Society Series B 31(3), 472–476.
- Draper, N. and W. Hunter (1969). Transformations: Some examples revisited. Technometrics 11(1), 23–40.
- Draper, N. R. and J. John (1981). Influential observations and outliers in regression. Technometrics 23(1), 21–26.
- Duan, N. (1993). Sensitivity analysis for Box-Cox power transformation model: Contrast parameters. Biometrika 80(4), 885–897.

- Durbin, B. P., J. S. Hardin, D. M. Hawkins, and D. M. Rocke (2002). A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 18(1), 105–110.
- Edgeworth, F. Y. (1886). XLVI. Problems in probabilities. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 22(137), 371–384.
- Edgeworth, F. Y. (1900). On the representation of statistics by mathematical formulae (supplement). Journal of the Royal Statistical Society 63(1), 72–81.
- Eicker, F. (1967). Limit theorems for regression with unequal and dependent errors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, 59–82.
- Eisele, M. and J. Zhu (2013). Multiple imputation in a complex household survey - the German Panel on Household Finances (PHF): Challenges and solutions. User Guide, Deutsche Bundesbank.
- Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics 3(1), 1–21.
- Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. Econometrica 71(1), 355–364.
- Elston, R. C. (1961). On additivity in the analysis of variance. Biometrics 17(2), 209–219.
- Emerson, J. D. and M. A. Stoto (1982). Exploratory methods for choosing power transformations. Journal of the American Statistical Association 77(377), 103–108.
- Empirica ag (2017). Regionaldatenbank Immobilien. <https://www.empirica-institut.de/>. [accessed: 05.10.2017].
- Erickson, B. H. and T. A. Nosanchuk (1977). Understanding Data. London: Open University Press.
- Eubank, R. L. (2004). Quantiles. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, and N. L. Johnson (Eds.), Encyclopedia of Statistical Sciences. Hoboken: John Wiley & Sons.
- EURAREA Consortium (2004). Enhancing small area estimation techniques to meet European needs. Project Reference Volume, Deliverable 7.1.4, Office of National Statistics.
- eurostat (2004). Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. EU-SILC 131-rev/04, Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, eurostat.
- eurostat (2013a). Handbook on precision requirements and variance estimation for ESS households survey. Methodologies and Working Papers, European Union.
- eurostat (2013b). Statistik der Europäischen Union über Einkommen und Lebensbedingungen (EU-SILC). <https://ec.europa.eu/eurostat/de/web/microdata/european-union-statistics-on-income-and-living-conditions>. [accessed: 18.09.2018].

- eurostat (2017). Europe 2020 indicators - poverty and social exclusion. http://ec.europa.eu/eurostat/statistics-explained/index.php/Europe_2020_indicators_-_poverty_and_social_exclusion. [accessed: 17.10.2017].
- eurostat (2018a). Distribution of income by quantiles - EU-SILC survey. http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di01&lang=en. [accessed: 12.04.2018].
- eurostat (2018b). Smarter, Greener, more Inclusive? Indicators to Support the Europe 2020 Strategy. Luxembourg: Publications Office of the European Union.
- Eurosystem Household Finance and Consumption Network (2013a). The Eurosystem Household Finance and Consumption Survey - methodological report for the first wave. Statistics Paper Series, European Central Bank.
- Eurosystem Household Finance and Consumption Network (2013b). The Eurosystem Household Finance and Consumption Survey - results from the first wave. Statistics Paper Series, European Central Bank.
- Fan, J., M. Tang, and M. Tian (2014). Kernel quantile estimator with ICI adaptive bandwidth selection technique. Acta Mathematica Sinica, English Series 30(4), 710–722.
- Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association 74(366), 269–277.
- Feng, C., H. Wang, N. Lu, T. Chen, H. He, Y. Lu, and X. Tu (2014). Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry 26(2), 105–109.
- Feng, Q., J. Hannig, and J. S. Marron (2016). A note on automatic data transformation. Stat 5(1), 82–87.
- Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. Biometrika 98(4), 995–999.
- Fernandez, E. S. (2014). Johnson: Johnson Transformation. R package version 1.4.
- Fink, E. L. (2009). The FAQs on data transformation. Communication Monographs 76(4), 379–397.
- Finney, D. (1941). On the distribution of a variate whose logarithm is normally distributed. Supplement to the Journal of the Royal Statistical Society 7(2), 155–161.
- Fisher, R. A. (1922a). On the interpretation of χ^2 from contingency tables, and the calculation of p. Journal of the Royal Statistical Society 85(1), 87–94.
- Fisher, R. A. (1922b). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London Series A 222(594-604), 309–368.

- Fisher, R. A. and F. Yates (1949). Statistical Tables for Biological, Agricultural and Medical Research. Edinburgh: Oliver and Boyd.
- Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: Wild bootstrap vs. pairs bootstrap. Computational Statistics & Data Analysis 49(2), 361–376.
- Fletcher, D., D. MacKenzie, and E. Villouta (2005). Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. Environmental and Ecological Statistics 12(1), 45–54.
- Forbes, C., M. Evans, N. Hastings, and B. Peacock (2011). Statistical Distributions. Hoboken: John Wiley & Sons.
- Forschungsdaten- und Servicezentrum (FDSZ) der Deutschen Bundesbank (2014). Panel on Household Finances (PHF). doi: 10.12757/Bbk.PHF.02.02.01. Plus one additional attribute (district code).
- Foster, A., L. Tian, and L. Wei (2001). Estimation for the Box-Cox transformation model without assuming parametric error distribution. Journal of the American Statistical Association 96(455), 1097–1101.
- Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica 52(3), 761–766.
- Fox, J. (1997). Applied Regression Analysis, Linear Models, and Related Methods. Thousand Oaks: SAGE Publications.
- Fox, J. and S. Weisberg (2011). An R Companion to Applied Regression. Thousand Oaks: SAGE Publications.
- Freeman, M. F. and J. W. Tukey (1950). Transformations related to the angular and the square root. The Annals of Mathematical Statistics 21(4), 607–611.
- Frick, J. R. and M. Grabka (2009). Gestiegene Vermögensungleichheit in Deutschland. DIW Wochenbericht, Deutsches Institut für Wirtschaftsforschung.
- Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. American Journal of Political Science 26(4), 797–833.
- Galton, F. (1889). Natural Inheritance. New York: Macmillan.
- Garson, G. D. (2012). Testing Statistical Assumptions. Asheboro: Statistical Associates Publishing.
- Gaudard, M. and M. Karson (2007). On estimating the Box-Cox transformation to normality. Communications in Statistics - Simulation and Computation 29(2), 559–582.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). Bayesian Data Analysis. Boca Raton: CRC Press.

- Gemmill, G. et al. (1980). Using the Box-Cox form for estimating demand: A comment. The Review of Economics and Statistics 62(1), 147–148.
- Genton, M. G., Y. Ma, and E. Parzen (2006). Discussion of "Sur une limitation très générale de la dispersion de la médiane" by M. Fréchet. Journal de la Société française de Statistique 147(2), 51–60.
- George, F. (2007). Johnson's System of Distributions and Microarray Data Analysis. Ph. D. thesis, University of South Florida.
- Geraci, M. (2016). **Qtools**: A collection of models and tools for quantile inference. The R Journal 8(2), 117–138.
- Gini, C. (1912). Variabilità e Mutabilità. Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche. Bologna: P. Cuppini.
- Goebel, J., M. M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp (2018). The German Socio-Economic Panel (SOEP). Journal of Economics and Statistics in print.
- Gómez-Rubio, V., N. Best, S. Richardson, G. Li, and P. Clarke (2010). Bayesian statistics for small area estimation. Technical report, Imperial College London.
- Gómez-Rubio, V., N. Salvati, et al. (2008). SAE2: Small Area Estimation with R. R package version 0.09.
- Goncalves, S. and N. Meddahi (2011). Box-Cox transforms for realized volatility. Journal of Econometrics 160(1), 129–144.
- Gottardo, R. and A. Raftery (2009). Bayesian robust transformation and variable selection: A unified approach. The Canadian Journal of Statistics 37(3), 361–380.
- Graf, M. and D. Nedyalkova (2014). Modeling of income and indicators of poverty and social exclusion using the generalized beta distribution of the second kind. The Review of Income and Wealth 60(4), 821–842.
- Gumbel, E. J. (1939). La probabilité des hypothèses. Comptes Rendus de l'Académie des Sciences 209, 645–647.
- Gurka, M. J., L. J. Edwards, K. E. Muller, and L. L. Kupper (2006). Extending the Box-Cox transformation to the linear mixed model. Journal of the Royal Statistical Society Series A 169(2), 273–288.
- Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society Series B 54(3), 761–771.
- Hagenaars, A., K. de Vos, and M. Zaidi (1994). Poverty Statistics in the Late 1980s: Research Based on Mirco-Data. Luxembourg: Office for the Official Publications of the European Communities.

- Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). Robust Statistics: The Approach based on Influence Functions. Hoboken: John Wiley & Sons.
- Han, A. K. (1987). A non-parametric analysis of transformations. Journal of Econometrics 35(2-3), 191–209.
- Harrell, F. E. and C. Davis (1982). A new distribution-free quantile estimator. Biometrika 69(3), 635–640.
- Harrell Jr, F. E., with contributions from Charles Dupont, and many others. (2018). **Hmisc: Harrell Miscellaneous**. R package version 4.1-1.
- Hartley, H. O. (1950). The use of range in analysis of variance. Biometrika 37(3/4), 271–280.
- Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. Biometrika 61(2), 383–385.
- Hawkins, D. M. (1980). Identification of Outliers. London: Chapman & Hall.
- Hawkins, D. M., D. Bradu, and G. V. Kass (1984). Location of several outliers in multiple-regression data using elemental sets. Technometrics 26(3), 197–208.
- Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. Transactions of the American Society of Civil Engineers 77, 1539–1641.
- Heagerty, P. and S. Zeger (2000). Marginalized multilevel models and likelihood inference. Statistical Science 15(1), 1–26.
- Hernandez, F. and R. A. Johnson (1980). The large-sample behavior of transformations to normality. Journal of the American Statistical Association 75(372), 855–861.
- Hey, G. (1938). A new method of experimental sampling illustrated on certain non-normal populations. Biometrika 30(1-2), 68–80.
- Hill, B. (1963). The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic. Journal of the American Statistical Association 58(301), 72–84.
- Hinkley, D. (1975). On power transformations to symmetry. Biometrika 62(1), 101–111.
- Hinkley, D. (1977). On quick choice of power transformation. Journal of the Royal Statistical Society Series C 26(1), 67–69.
- Hinkley, D. (1985). Transformation diagnostics for linear models. Biometrika 72(3), 487–496.
- Hinkley, D. and G. Runger (1984). The analysis of transformed data. Journal of the American Statistical Association 79(386), 302–309.
- Hoaglin, D. C., F. Mosteller, and W. T. Tukey (2000). Understanding Robust and Exploratory Data Analysis. Hoboken: John Wiley & Sons.

- Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67.
- Hoeting, J. A. and J. G. Ibrahim (1998). Bayesian predictive simultaneous variable and transformation selection in the linear model. Computational Statistics & Data Analysis 28(1), 87–103.
- Hoeting, J. A., A. E. Raftery, and D. Madigan (2002). Bayesian variable and transformation selection in linear regression. Journal of Computational and Graphical Statistics 11(3), 485–507.
- Hosking, J. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal for the Royal Statistical Society Series B 52(1), 105–124.
- Hossain, M. Z. (2011). The use of Box-Cox transformation technique in economic and statistical analyses. Journal of Emerging Trends in Economics and Management Sciences 2(1), 32–39.
- Household Finance and Consumption Network (2016a). The Household Finance and Consumption Survey: Methodological report for the second wave. Statistics Paper Series, European Central Bank.
- Household Finance and Consumption Network (2016b). The Household Finance and Consumption Survey: Results from the second wave. Statistics Paper Series, European Central Bank.
- Hoyle, M. H. (1973). Transformations: An introduction and a bibliography. International Statistical Review 41(2), 203–223.
- Huang, E. T. and W. R. Bell (2012). An empirical study on using previous American Community Survey data versus census 2000 data in SAIPE models for poverty estimates. Research Report Series, U.S. Census Bureau.
- Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistical 35(1), 73–101.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1, 221–233.
- Huber, P. J. (1981). Robust Statistics. Hoboken: John Wiley & Sons.
- Huber, P. J. (1992). Robust estimation of a location parameter. In S. Kotz and N. L. Johnson (Eds.), Breakthroughs in Statistics, pp. 492–518. New York: Springer-Verlag.
- Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2003). Parameter estimation for the calibration and variance stabilization of microarray data. Statistical Applications in Genetics and Molecular Biology 2(1), 1–24.

- Hulten, C. R. and F. C. Wykoff (1981). The estimation of economic depreciation using vintage asset prices: An application of the Box-Cox power transformation. Journal of Econometrics 15(3), 367–396.
- Hutcheson, G. and N. Sofroniou (1999). The Multivariate Social Scientist: Introductory Statistics using Generalized Linear Models. Thousand Oaks: SAGE Publications.
- Hyndman, R. and Y. Fan (1996). Sample quantiles in statistical packages. The American Statistician 50(4), 361–365.
- IBM Corp (2013). IBM SPSS Statistics for Windows, Version 25.0.
- Jeffreys, H. (1998). The Theory of Probability. Oxford: Oxford University Press.
- Jensen, D. R. and H. Solomon (1972). A Gaussian approximation to the distribution of a definite quadratic form. Journal of the American Statistical Association 67(340), 898–902.
- Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. Acta Mathematica 30, 175–193.
- John, J. A. and N. R. Draper (1980). An alternative family of transformations. Journal of the Royal Statistical Society Series C 29(2), 190–197.
- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. Biometrika 36(1/2), 149–176.
- Johnson, N. L. and S. Kotz (1970). Continuous Univariate Distributions. Boston: Houghton Mifflin Harcourt.
- Johnson, R. A. (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79(386), 314–315.
- Johnson, R. A. (2009). Statistics - Principles and Methods. Hoboken: John Wiley & Sons.
- Johnston, J. and J. DiNardo (1972). Econometric Methods. New York: McGraw-Hill Education.
- Jones, M. C. and A. Pewsey (2009). Sinh-arcsinh distributions. Biometrika 96(4), 761–780.
- Juritz, J. M., J. W. F. Juritz, and M. Stephens (1983). On the accuracy of simulated percentage points. Journal of the American Statistical Association 78(382), 441–444.
- Kavonius, I. and J. Honkkila (2013). Reconciling micro and macro data on household wealth: A test based on three euro area countries. Journal of Economic and Social Policy 15(2), 1–30.
- Keene, O. N. (1995). The log transformation is special. Statistics in Medicine 14(8), 811–819.
- Kelmansky, D. M., E. J. Martínez, and V. Leiva (2013). A new variance stabilizing transformation for gene expression data analysis. Statistical Applications in Genetics and Molecular Biology 12(6), 653–666.

- Kelmansky, D. M. and L. Ricci (2017). A new distribution family for microarray data. Microarrays 6(1), 1–5.
- Kendall, M. G. (1938). A new measure of rank correlation. Biometrika 30(1/2), 81–93.
- Kettl, S. (1991). Accounting for heteroscedasticity in the transform both sides regression model. Journal of the Royal Statistical Society Series C 40(2), 261–268.
- Kim, C., B. E. Storer, and M. Jeong (1996). Note on Box-Cox transformation diagnostics. Technometrics 38(2), 178–180.
- Kim, J., J. Brick, W. Fuller, and G. Kalton (2006). On the bias of the multiple-imputation variance estimator in survey sampling. Journal of the Royal Statistical Society Series B 68(3), 509–521.
- Kirk, R. E. (1968). Experimental Design: Procedures for the Behavioral Sciences. Thousand Oaks: SAGE Publications.
- Kleczkowski, A. (1949). The transformation of local lesion counts for statistical analysis. Annals of Applied Biology 36(1), 139–152.
- Kleiber, C. and S. Kotz (2003). Statistical Size Distributions in Economics and Actuarial Sciences. Hoboken: John Wiley & Sons.
- Knerr, P., F. Aust, N. Chudziak, R. Gilberg, and M. Kleudgen (2015). Methodenbericht - Private Haushalte und ihre Finanzen (PHF) 2. Erhebungswelle - Anonymisierte Fassung -. Methodenbericht, infas Institut für angewandte Sozialwissenschaft GmbH.
- Koenker, R. (2005). Quantile Regression. Cambridge: Cambridge University Press.
- Kolenikov, S. (2017). epctile - Estimation and inference for percentiles.
- Kott, P. (1995). A paradox of multile imputation. Proceedings Survey Research Methods Section, American Statistical Association.
- Krasker, W. S. and R. E. Welsch (1982). Efficient bounded-influence regression estimation. Journal of the American Statistical Association 77(379), 595–604.
- Kreutzmann, A.-K. (2016). Poverty mapping using small area estimation: An application in R. Master's thesis, Freie Universität Berlin.
- Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2018). emdi: Estimating and Mapping Disaggregated Indicators. R package version 1.1.3.
- Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2019). The R package **emdi** for estimating and mapping regionally disaggregated indicators. Journal of Statistical Software. forthcoming.
- Kruskal, J. B. (1968). Statistical analysis: Transformations of data. In D. L. Sills and R. K. Merton (Eds.), International Encyclopedia of the Social Sciences, pp. 182–193. New York: Macmillan Publishers.

- Kruskal, W. H. and W. A. Wallis (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47(260), 583–621.
- Kuhn, M. (2008). Building predictive models in R using the **caret** package. Journal of Statistical Software 28(5), 1–26.
- Kullback, S. (1997). Information Theory and Statistics. Mineola: Dover Publications.
- Kumhof, M., R. Ranci ere, and P. Winant (2015). Inequality, leverage, and crises. American Economic Review 105(3), 1217–1245.
- Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. SIAM Journal on Optimization 9(1), 112–147.
- Lahiri, P. and J. Suntornc host (2015). Variable selection for linear mixed models with application in small area estimation. The Indian Journal of Statistics 77(2), 312–320.
- Laird, M. N. and J. H. Ware (1983). Random-effects models for longitudinal data. Biometrics 38(4), 963–74.
- Landesamt f ur Statistik Niedersachsen (2014). Geb aude- und Wohnungsbestand in Deutschland - Erste Ergebnisse der Geb aude- und Wohnungsz ahlung 2011. Technical report, Statistische  mter des Bundes und der L ander.
- Langford, E. (2006). Quartiles in elementary statistics. Journal of Statistics Education 50(4), 361–365.
- Laubscher, N. F. (1961). On stabilizing the binomial and negative binomial variances. Journal of the American Statistical Association 56(293), 143–150.
- Laud, P. W. and J. G. Ibrahim (1995). Predictive model selection. Journal of the Royal Statistical Society Series B 57(1), 247–262.
- Lavall e, P. and J.-F. Beaumont (2015). Why we should put some weight on weights. Survey Methods: Insights from the Field, 1–18.
- Lawrance, A. (1987a). A note on the variance of the Box-Cox regression transformation estimate. Journal of the Royal Statistical Society Series C 36(2), 221–223.
- Lawrance, A. (1987b). The score statistic for regression transformation. Biometrika 74(2), 275–379.
- Leadership Council of the Sustainable Development Solutions Network (2015). Indicators and a monitoring framework for the Sustainable Development Goals. Report to the Secretary-General of the United Nations, United Nations.
- L’Ecuyer, P. (1999). Good parameters and implementations for combined multiple recursive random number generators. Operations Research 47(1), 159–164.

- L'Ecuyer, P., R. Simard, E. J. Chen, and W. D. Kelton (2002). An object-oriented random-number package with many long streams and substreams. Operations Research 50(6), 1073–1075.
- Lee, C. (1982). Comparison of two correction methods for the bias due to the logarithmic transformation in the estimation of biomass. Canadian Journal of Forest Research 12(2), 326–331.
- Lee, Y., J. A. Nelder, et al. (2004). Conditional and marginal models: Another view. Statistical Science 19(2), 219–238.
- Leinhardt, S. and S. S. Wasserman (1979). Exploratory data analysis: An introduction to selected methods. Sociological Methodology 10, 311–365.
- Lesaffre, E. and G. Molenberghs (1991). Multivariate probit analysis: A neglected procedure in medical statistics. Statistics in Medicine 10(9), 1391–1403.
- Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, pp. 278–292. Palo Alto: Stanford University Press.
- Leydold, J. (2017). **rstream**: Streams of Random Numbers. R package version 1.3.5.
- Li, H. and P. Lahiri (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis 101(4), 882–892.
- Lipsitz, S. R., J. Ibrahim, and G. Molenberghs (2000). Using a Box-Cox transformation in the analysis of longitudinal data with incomplete responses. Journal of the Royal Statistical Society Series C 49(3), 287–296.
- Litière, S., A. Alonso, and G. Molenberghs (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Statistics in Medicine 27(16), 3125–3144.
- Lo, S. and S. Andrews (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. Frontiers in Psychology 6(1171), 1–16.
- Lohr, S. L. (2010). Sampling: Design and Analysis. Boston: Cengage Learning.
- Longford, N. (2011). Small-sample estimators of the quantiles of the normal, log-normal and pareto distributions. Journal of the Statistical Computation and Simulation 82(9), 1383–1395.
- Lopez-Vizcaino, E., M. Lombardia, and D. Morales (2014). **mme**: Multinomial Mixed Effects Models. R package version 0.1-5.
- Lumley, T. (2004). Analysis of complex survey samples. Journal of Statistical Software 9(8), 1–19.

- Ma, Y., M. G. Genton, and E. Parzen (2011). Asymptotic properties of sample quantiles of discrete distributions. Annals of the Institute of Statistical Mathematics 63(2), 227–243.
- Machado, J. A. F. and J. Mata (2000). Box-Cox quantile regression and the distribution of firm sizes. Journal of Applied Econometrics 15(3), 253–274.
- MacKinnon, J. G. and L. Magee (1990). Transforming the dependent variable in regression models. International Economic Review 31(2), 315–339.
- Magee, L. (1988). The behaviour of a modified Box-Cox regression model when some values of the dependent variable are close to zero. The Review of Economics and Statistics 70(2), 362–366.
- Majumder, K. L. and G. P. Bhattacharjee (1973). Algorithm AS63: The incomplete beta integral. Journal of the Royal Statistical Society Series C 22(3), 409–411.
- Makkonen, L. and M. Pajari (2014). Defining sample quantiles by the true rank probability. Journal of Probability and Statistics, 1–6.
- Manly, B. F. J. (1976). Exponential data transformations. Journal of the Royal Statistical Society Series D 25(1), 37–42.
- Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17(3), 283–295.
- Marazzi, A. and V. J. Yohai (2004). Robust Box-Cox transformations for simple regression. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst (Eds.), Theory and Applications of Recent Robust Methods, pp. 173–182. Basel: Birkhäuser Basel.
- Marazzi, A. and V. J. Yohai (2006). Robust Box-Cox transformations based on minimum residual autocorrelation. Computational Statistics & Data Analysis 50(10), 2752–2768.
- Marchetti, S., M. Beręsewicz, N. Salvati, M. Szymkowiak, and Ł. Wawrowski (2018). The use of a three-level M-quantile model to map poverty at local administrative unit 1 in Poland. Journal of the Royal Statistical Society Series A 181(4), 1–28.
- Marchetti, S., C. Giusti, and M. Pratesi (2016). The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy. AStA Wirtschafts- und Sozialstatistisches Archiv 10(2-3), 79–93.
- Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli (2015). Small area model-based estimators using big data sources. Journal of Official Statistics 31(2), 263–281.
- Marhuenda, Y., D. Morales, and M. Pardo (2014). Information criteria for Fay-Herriot model selection. Computational Statistics and Data Analysis 70, 268–280.
- Marshall, A., D. Altman, R. Holder, and P. Royston (2009). Combining estimates of interest in prognostic modelling studies after multiple imputation: Current practice and guidelines. BMC Medical Research Methodology 9(57), 1–8.

- Maruo, K., Y. Yamaguchi, H. Noma, and M. Goshō (2017). Interpretable inference on the mixed effect model with the Box-Cox transformation. Statistics in Medicine 36(15), 2420–2434.
- McCulloch, C. E. and J. M. Neuhaus (2001). Generalized Linear Mixed Models. Hoboken: John Wiley & Sons.
- McDonald, J. (1984). Some generalized functions for the size distribution of income. Econometrica 52(3), 647–663.
- McDonald, J. and R. Bordley (1996). Something new, something old: Parametric models for the size distribution of income. Journal of Income Distribution 6(1), 91–103.
- McNeil, D. R. (1977). Interactive Data Analysis: A Practical Primer. Hoboken: John Wiley & Sons.
- Medina, L. (2017). Transformations in the linear regression model: An overview. Master's thesis, Freie Universität Berlin.
- Medina, L., P. Castro, A.-K. Kreuzmann, and N. Rojas-Perilla (2018). **trafo**: Estimation, Comparison and Selection of Transformations. R package version 1.0.0.
- Miller, J. P. (2010). Essential Statistical Methods for Medical Statistics. Amsterdam: Elsevier.
- Mills, T. C. (1978). The functional form of the U.K. demand for money. Journal of the Royal Statistical Society Series C 27(1), 52–57.
- Molina, I. and Y. Marhuenda (2015). **sae**: An R package for small area estimation. The R Journal 7(1), 81–98.
- Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics 38(3), 369–385.
- Montgomery, D. C. (2008). Design and Analysis of Experiments. Hoboken: John Wiley & Sons.
- Moore, P. G. (1957). Transformations to normality using fractional powers of the variable. Journal of the American Statistical Association 52(278), 237–246.
- Moore, P. G. (1958). Interval analysis and the logarithmic transformation. Journal of the Royal Statistical Society Series B 20(1), 187–192.
- Moore, P. G. and J. W. Tukey (1954). Answer to query 112. Biometrics 10(4), 562–568.
- Morozova, M., K. Koschutnig, E. Klein, and G. Wood (2016). Monotonic non-linear transformations as a tool to investigate age-related effects on brain white matter integrity: A Box-Cox investigation. NeuroImage 125, 1119–1130.
- Moschopoulos, P. G. (1983). On a new transformation to normality. Communications in Statistics - Theory and Methods 12(16), 1873–1878.

- Mosteller, F. and R. R. Bush (1954). Selected quantitative techniques. In G. Lindzey (Ed.), Handbook of Social Psychology. Boston: Addison-Wesley.
- Mosteller, F. and J. W. Tukey (1977). Data Analysis and Regression: A Second Course in Statistics. Boston: Addison-Wesley.
- Mosteller, F. and C. Youtz (2006). Tables of the Freeman-Tukey transformations for the binomial and poisson distributions. In S. E. Fienberg and D. C. Hoaglin (Eds.), Selected Papers of Frederick Mosteller, pp. 337–347. New York: Springer-Verlag.
- Moura, F., A. Neves, and D. do N. Silva (2017). Small area models for skewed Brazilian business survey data. Journal of the Royal Statistical Society Series A 180(4), 1039–1055.
- Muenchen, R. A. (2017). The popularity of data science software. <http://r4stats.com/articles/popularity/>. [accessed: 27.02.2018].
- Mukhopadhyay, P. K. and A. McDowell (2011). Small area estimation for survey data analysis using SAS software. SAS Conference Proceedings, SAS Institute.
- Müller, S., J. L. Scealy, A. H. Welsh, et al. (2013). Model selection in linear mixed models. Statistical Science 28(2), 135–167.
- Münnich, R., J. P. Burgard, and M. Vogt (2013). Small Area-Statistik: Methoden und Anwendungen. AStA Wirtschafts- und Sozialstatistisches Archiv 6(3-4), 149–191.
- Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2), 133–142.
- Natrella, M. G. (2013). Experimental Statistics. North Chelmsford: Courier Corporation.
- Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. The Computer Journal 7(4), 308–313.
- Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. Journal of the Royal Statistical Society Series A 135(3), 370–384.
- Neves, A., D. Silva, and S. Correa (2013). Small domain estimation for the Brazilian service sector survey. ESTADÍSTICA 65(185), 13–37.
- Newman, M. C. (1993). Regression analysis of log-transformed data: Statistical bias and its correction. Environmental Toxicology and Chemistry 12(6), 1129–1133.
- Neyman, J. and E. L. Scott (1960). Correction for bias introduced by a transformation of variables. The Annals of Mathematical Statistics 31(3), 643–655.
- Nychka, D. and D. Ruppert (1995). Nonparametric transformations for both sides of a regression model. Journal of the Royal Statistical Society Series B 57(3), 519–532.
- O’Hara, R. B. and D. J. Kotze (2010). Do not log-transform count data. Methods in Ecology and Evolution 1(2), 118–122.

- Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. Scandinavian Journal of Statistics 8(3), 154–168.
- Okolewski, A. and T. Rychlik (2001). Sharp distribution-free bounds on the bias in estimating quantiles via order statistics. Statistics & Probability Letters 52(2), 207–213.
- Osborne, J. W. (2002). The effects of minimum values on data transformations. Practical Assessment, Research & Evaluation 8(6), 1–7.
- Osborne, J. W. and A. Overbay (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research, & Evaluation 9(6), 1–8.
- Osborne, J. W. and E. Waters (2012). Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research, & Evaluation 8(2), 1–5.
- Ottaviano, G. I. P. and D. Puga (1998). Agglomeration in the global economy: A survey of the 'new economic geography'. World Economy 21(6), 707–731.
- Parrish, R. (1990). Comparison of quantile estimators in normal sampling. Biometrics 46(1), 247–257.
- Parzen, E. (1979). Nonparametric statistical data modeling. Journal of the American Statistical Association 74(365), 105–121.
- Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. Biometrika 23(1/2), 114–133.
- Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London Series A 185, 71–110.
- Pebesma, E. (2018). sf: Simple Features for R. R package version 0.6-0.
- Peng Zhang, Peter X.-K. Song, A. Q. and T. Greene (2008). Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. Biometrics 64(1), 29–38.
- Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. Biometrika 68(1), 35–43.
- Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science 28(1), 40–68.
- Phien, H. (1990). A note on the computation of the incomplete beta function. Advances in Engineering Software 12(1), 39–44.
- Piacentini, M. (2014). Measuring income inequality and poverty at the regional level in OECD countries. OECD Statistics Working Papers, Organisation for Economic Co-operation and Development.

- Piepho, H.-P. and C. E. McCulloch (2004). Transformations in mixed models: Application to risk analysis for a multienvironment trial. Journal of Agricultural, Biological, and Environmental Statistics 9(2), 123–137.
- Piketty, T. and G. Zucman (2014). Capital is back: Wealth-income ratios in rich countries 1700-2010. The Quarterly Journal of Economics 129(3), 1255–1310.
- Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team (2015). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-122.
- Poirier, D. J. (1978). The use of the Box-Cox transformation in limited dependent variable models. Journal of the American Statistical Association 73(362), 284–287.
- Powers, D., W. Basel, and B. O’Hara (2008). SAIPE county poverty models using data from the American Community Survey. Technical report, U.S. Census Bureau.
- Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small-area estimation. Journal of the American Statistical Association 85(409), 163–171.
- Pratesi, M. and N. Salvati (2009). Small area estimation in the presence of correlated random area effects. Journal of Official Statistics 25(1), 37–53.
- Rabe-Hesketh, S. and A. Skrondal (2012). Multilevel and Longitudinal Modeling Using Stata. College Station: Stata Press.
- Rahman, M. (1999). Estimating the Box-Cox transformation via Shapiro-Wilk W statistic. Communications in Statistics - Simulation and Computation 28(1), 223–241.
- Rahman, M. and L. Pearson (2008). Anderson-Darling statistic in estimating the Box-Cox. Journal of Applied Probability & Statistics 3(1), 45–57.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society Series B 31(2), 350–371.
- Ramsey, J. B. (1974). Classical model selection through specification error tests. In P. Zarembka (Ed.), Frontiers in Econometrics. New York: Academic Press.
- Rao, J. N. K. (1999). Some recent advances in model-based small area estimation. Survey Methodology 25(2), 175–186.
- Rao, J. N. K. and I. Molina (2015). Small Area Estimation. Hoboken: John Wiley & Sons.
- Rao, J. N. K. and C. F. J. Wu (1988). Resampling inference with complex survey data. Journal of the American Statistical Association 83(401), 231–241.
- Raudenbush, S. and A. Bryk (2002). Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks: SAGE Publications.
- R Core Team (2018). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

- Rendtel, U. (1995). Lebenslagen im Wandel: Panelfälle und Panelrepräsentativität. Frankfurt: Campus Verlag.
- Ribeiro Jr., P. J. and P. J. Diggle (2016). **geoR**: Analysis of geostatistical data. R News 1(2), 15–18.
- Rocke, D. M. (1993). On the beta transformation family. Technometrics 35(1), 72–81.
- Rodríguez-Pose, A. and R. Crescenzi (2008). Mountains in a flat world: Why proximity still matters for the location of economic activity. Cambridge Journal of Regions, Economy and Society 1(3), 371–388.
- Rojas-Perilla, N. (2018). The Use of Data-Driven Transformations and their Application in Small Area Estimation. Ph. D. thesis, Freie Universität Berlin. Unpublished.
- Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2017). Data-driven transformations in small area estimation. Discussion Paper, School of Business and Economics.
- Rosenthal, J. A. (2011). Statistics and Data Interpretation for Social Work. New York: Springer Publishing Company.
- Rothery, P. (1988). A cautionary note on data transformation: Bias in back-transformed means. Bird Study 35(3), 219–221.
- Rousseeuw, P. J. and A. M. Leroy (2005). Robust Regression and Outlier Detection. Hoboken: John Wiley & Sons.
- Rousseeuw, P. J. and B. C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85(41), 633–639.
- Royston, P., P. C. Lambert, et al. (2011). Flexible Parametric Survival Analysis using Stata: Beyond the Cox Model. College Station: Stata Press.
- Rubin, D. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Society 91(434), 473–489.
- Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3), 581–592.
- Rubin, D. B. (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79(386), 309–312.
- Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Hoboken: John Wiley & Sons.
- Ruppert, D. (2001). Statistical analysis, special problems of: Transformations of data. In N. J. Smelser and P. B. Baltes (Eds.), The International Encyclopedia of the Social & Behavioral Sciences, pp. 15007–15014. Amsterdam: Elsevier.
- Ruppert, D. and B. Aldershof (1989). Transformations to symmetry and homoscedasticity. Journal of the American Statistical Association 84(406), 437–446.

- Rust, K. F. and J. N. K. Rao (1996). Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research 5(3), 283–310.
- RWI;microm (2016a). Socio-Economic Data on grid level. <http://fdz.rwi-essen.de/microm.html>. [accessed: 08.06.2018].
- RWI;microm (2016b). Socio-Economic Data on grid level. Car segments. <http://doi.org/10.7807/microm:pkwseg:v4>. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.
- RWI;microm (2016c). Socio-Economic Data on grid level. House typ. <http://doi.org/10.7807/microm:haustyp:v4>. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.
- RWI;microm (2016d). Socio-Economic Data on grid level. Household structure. <http://doi.org/10.7807/microm:hstruktur:v4>. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.
- RWI;microm (2016e). Socio-Economic Data on grid level. Payment index. <http://doi.org/10.7807/microm:zahlindex:v4>. RWI-GEO-GRID. Version: 1. RWI - Leibniz-Institut für Wirtschaftsforschung. Datensatz.
- Sakia, R. M. (1990). Retransformation bias: A look at the Box-Cox transformation to linear balanced mixed ANOVA models. Metrika 37(1), 345–351.
- Sakia, R. M. (1992). The Box-Cox transformation technique: A review. Journal of the Royal Statistical Society Series D 41(2), 169–178.
- SAS Institute Inc. (2018). Version 9.4 of the SAS System.
- Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes using mobile phone data: Estimating literacy rates in Senegal. Journal of the Royal Statistical Society Series A 180(4), 1163–1190.
- Schmid, T., N. Tzavidis, R. Münnich, and R. Chambers (2016). Outlier robust small area estimation under spatial correlation. Scandinavian Journal of Statistics 43(3), 806–826.
- Schoch, T. (2014). rsae: Robust Small Area Estimation. R package version 0.1-5.
- Schoonjans, F., D. De Bacquer, and P. Schmid (2011). Estimation of population percentiles. Epidemiology 22(5), 750–751.
- Sfakianakis, M. and D. Verginis (2008). A new family of nonparametric quantile estimators. Communications in Statistics - Simulation and Computation 37(2), 337–345.
- Shao, J. (1988). A note on bootstrap variance estimation. Technical report, Purdue University.
- Shao, J. and C. Wu (1989). A general theory for jackknife variance estimation. The Annals of Statistics 17(3), 1176–1197.

- Shao, J. and C. Wu (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. The Annals of Statistics 20(3), 1571–1593.
- Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality. Biometrika 52(3/4), 591–611.
- Sheather, S. and J. Marron (1990). Kernel quantile estimators. Journal of the American Statistical Association 85(410), 410–416.
- Shi, C. and with contributions from Peng Zhang (2013). **BayesSAE: Bayesian Analysis of Small Area Estimation**. R package version 1.0-1.
- Shin, Y. (2008). Semiparametric estimation of the Box-Cox transformation model. The Econometrics Journal 11(3), 517–537.
- Slifker, J. F. and S. S. Shapiro (1980). The Johnson system: Selection and parameter estimation. Technometrics 22(2), 239–246.
- Slud, E. and T. Maiti (2006). Mean-squared error estimation in transformed Fay-Herriot models. Journal of the Royal Statistical Society Series B 68(2), 239–257.
- Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19(2), 279–281.
- Snedecor, G. W. and W. G. Cochran (1989). Statistical Methods. Ames: Iowa State University Press.
- Sokal, R. R. and F. J. Rohlf (1995). Biometry: The Principles and Practice of Statistics in Biological Research. New York: W.H. Freeman and Company.
- Solomon, P. J. (1985). Transformations for components of variance and covariance. Biometrika 72(2), 233–239.
- Spanos, A. (1986). Statistical Foundations of Econometric Modelling. Cambridge: Cambridge University Press.
- StataCorp (2015). Stata Statistical Software: Release 15. College Station: StataCorp LLC.
- StataCorp (2017). Stata multilevel mixed-effects. Reference Manual, Stata Press.
- Statistics Canada (2013). 2013 National Graduate Survey (class of 2009-2010). Microdata User Guide, Statistics Canada.
- Statistik Austria (2013). Registerbasierte Statistiken Demographie (RS). Schnellbericht 10.7.
- Statistische Ämter des Bundes und der Länder (2011). Erwerbstätige Bevölkerung im regionalen Vergleich nach Stellung im Beruf. https://ergebnisse.zensus2011.de/#StaticContent:00,BEG_4_3_2,,https://ergebnisse.zensus2011.de/#StaticContent:00,BEG_4_3_2,,. Zensus 2011 [accessed: 05.06.2018].

- Statistische Ämter des Bundes und der Länder (2011). Zensus 2011. <https://ergebnisse.zensus2011.de>. [accessed: 05.06.2018].
- Statistische Ämter des Bundes und der Länder (2014a). Arbeitslose nach ausgewählten Personengruppen sowie Arbeitslosenquoten - Jahresdurchschnitt - regionale Ebenen. <https://www.regionalstatistik.de/genesis/online/data;jsessionid=303A27704A8955EAFF3BEDB689D51244.reg2?operation=abruftabelleBearbeiten&levelindex=2&levelid=1528188166412&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordnungsstruktur&auswahlziel=werteabruf&selectionname=13211-02-05-4-B&auswahltext=&nummer=10&variable=10&name=DLAND&werteabruf=Werteabruf>. Regionaldatenbank Deutschland [accessed 05.06.2018].
- Statistische Ämter des Bundes und der Länder (2014b). Indikatoren des Indikatorensystems "Nachhaltigkeit" Themenbereich "Bevölkerung". <https://www-genesis.destatis.de/gis/genView?GenMLURL=https://www-genesis.destatis.de/regatlas/AI-N-04.xml&CONTEXT=REGATLAS01>. Regionalatlas Deutschland [accessed: 05.06.2018].
- Statistische Ämter des Bundes und der Länder (2014c). Indikatoren des Themenbereichs "Bevölkerung". <https://www-genesis.destatis.de/gis/genView?GenMLURL=https://www-genesis.destatis.de/regatlas/AI002-2.xml&CONTEXT=REGATLAS01>. Regionalatlas Deutschland [accessed: 05.06.2017].
- Statistische Ämter des Bundes und der Länder (2014d). Sparen der privaten Haushalte 1991 bis 2012 (WZ2008). <https://www.statistik-bw.de/VGRdL/tbls/tab.jsp?rev=RV2011&tbl=tab15&lang=de-DE#tab05>. Volkswirtschaftliche Gesamtrechnungen der Länder VGRdL [accessed: 05.06.2018].
- Statistische Ämter des Bundes und der Länder (2014e). Verfügbares Einkommen 1991 bis 2016 (WZ2008). <https://www.statistik-bw.de/VGRdL/tbls/tab.jsp?rev=RV2014&tbl=tab14&lang=de-DE#tab05>. Volkswirtschaftliche Gesamtrechnungen der Länder VGRdL [accessed: 05.06.2018].
- Statistische Ämter des Bundes und der Länder (2018). Gemeinsames Statistikportal. <https://www.statistikportal.de/de/node/150>. [accessed: 05.06.2018].
- Steinhauer, H. W., C. Abmann, S. Zinn, S. Goßmann, and S. Rässler (2015). Sampling and weighting cohort samples in institutional contexts. ASTA Wirtschafts- und Sozialstatistisches Archiv 9(2), 131–157.
- Steorts, R. C. and M. Ghosh (2013). On estimation of mean squared errors of benchmarked empirical bayes estimators. Statistica Sinica 23(2), 749–767.
- Sugawasa, S. and T. Kubokawa (2017). Transforming response values in small area prediction. Computational Statistics and Data Analysis 114, 47–60.

- Sweeting, T. J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model. Biometrika 71(1), 127–134.
- Tabachnick, B. G. and L. S. Fidell (2007). Using Multivariate Statistics. London: Pearson.
- Taylor, J. M. G. (1985). Power transformations to symmetry. Biometrika 72(1), 145–152.
- Thai, H.-T., F. Mentré, N. H. Holford, C. Veyrat-Follet, and E. Comets (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. Pharmaceutical Statistics 12(3), 129–140.
- The World Bank (2007). More than a pretty picture: Using poverty maps to design better policies and interventions. Report, The International Bank for Reconstruction and Development/The World Bank.
- The World Bank (2017). Measuring poverty. <http://www.worldbank.org/en/topic/measuringpoverty>. [accessed: 27.04.2017].
- The World Bank Group (2013). Software for poverty mapping. <http://go.worldbank.org/QG9L6V7P20>. [accessed: 11.02.2016].
- Thoni, H. (1969). Transformation of variables used in the analysis of experimental and observational data: A review. Technical report, Iowa State University.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B 58(1), 267–288.
- Tierney, L., A. J. Rossini, N. Li, and H. Sevcikova (2016). snov: Simple Network of Workstations. R package version 0.4-2.
- Tsai, A. C., M. Liou, M. Simak, and P. E. Cheng (2017). On hyperbolic transformations to normality. Computational Statistics & Data Analysis 115, 250–266.
- Tsai, C.-L. and X. Wu (1990). Diagnostics in transformation and weighted regression. Technometrics 32(3), 315–322.
- Tukey, J. W. (1949). One degree of freedom for non-additivity. Biometrics 5(3), 232–242.
- Tukey, J. W. (1957). On the comparative anatomy of transformations. The Annals of Mathematical Statistics 28(3), 602–632.
- Tukey, J. W. (1977). Exploratory Data Analysis. London: Pearson.
- Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small area means and quantiles. Australian and New Zealand Journal of Statistics 52(2), 167–186.
- Tzavidis, N., L.-C. Zhang, A. Luna, T. Schmid, and N. Rojas-Perilla (2018). From start to finish: A framework for the production of small area official statistics. Journal of the Royal Statistical Society Series A 181(4), 927–979.

- Urbanek, S. (2009 - 2014). **multicore**: Parallel Processing of R Code on Machines with Multiple Cores or CPUs.
- Ushey, K., J. McPherson, J. Cheng, A. Atkins, and J. Allaire (2018). **packrat**: A Dependency Management System for Projects and their R Package Dependencies. R package version 0.4.9-1.
- Vélez, J. I. and J. C. Correa (2014). Should we think of a different median estimator? Comunicaciones en Estadística 7(1), 11–17.
- Vélez, J. I., J. C. Correa, and F. Marmolejo-Ramos (2015). A new approach to the Box-Cox transformation. Frontiers in Applied Mathematics and Statistics 1(12), 1–10.
- Velilla, S. (1993). Quantile-based estimation for the Box-Cox transformation in random samples. Statistics & Probability Letters 16(2), 137–145.
- Venables, W. N. and B. D. Ripley (2002). Modern Applied Statistics with S. New York: Springer-Verlag.
- Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association 91(433), 217–221.
- Verbeke, G. and G. Molenberghs (2000). Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.
- Walker, A. (2017). **openxlsx**: Read, Write and Edit XLSX Files. R package version 4.0.17.
- Walker, A. M. (1968). A note on the asymptotic distribution of sample quantiles. Journal of the Royal Statistical Society Series B 30(3), 570–575.
- Wang, N. and D. Ruppert (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. Journal of the American Statistical Association 90(430), 522–534.
- Wang, S. (1987). Improved approximation for transformation diagnostics. Communications in Statistics - Theory and Methods 16(6), 1797–1819.
- Wang, Y. (2015). **jtrans**: Johnson Transformation for Normality. R package version 1.1.0.
- Warnholz, S. (2016a). **saeRobust**: Robust Small Area Estimation. R package version 0.1.0.
- Warnholz, S. (2016b). Small Area Estimation using Robust Extensions to Area Level Models. Ph. D. thesis, Freie Universität Berlin.
- Warton, D. I. and F. K. C. Hui (2011). The arcsine is asinine: The analysis of proportions in ecology. Ecology 92(1), 3–10.
- Wattanacheewakul, L. (2014). A new family of transformations for lifetime data. Proceedings of the World Congress on Engineering, International Association of Engineers.

- Wei, L., D. Wang, and A. Hutson (2015). An investigation of quantile function estimators relative to quantile confidence interval coverage. Communications in Statistics - Theory and Methods 44(10), 2107–2135.
- Weibull, W. (1939). The phenomenon of rupture in solids. Ingeniörs Vetenskaps Akademien Handlingar 17(153), 1–55.
- Weisberg, S. (1980). Applied Linear Regression. Hoboken: John Wiley & Sons.
- West, B. T., K. B. Welch, and A. T. Galecki (2007). Linear Mixed Models: A Practical Guide Using Statistical Software. Boca Raton: CRC Press.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48(4), 817–838.
- Whittaker, J., C. Whitehead, and M. Somers (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. Journal of the Royal Statistical Society Series C 54(4), 863–878.
- Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.
- Wilcox, R. R. (2005). Introduction to Robust Estimation and Hypothesis Testing. New York: Elsevier.
- Williams, M., C. A. G. Grajales, and D. Kurkiewicz (2013). Assumptions of multiple regression: Correcting two misconceptions. Practical Assessment, Research & Evaluation 18(11), 1–14.
- Wilson, E., M. Underwood, O. Puckrin, K. Letto, R. Doyle, H. Caravan, S. Camus, and K. Bassett (2013). The arcsine transformation: Has the time come for retirement? Technical report.
- Wilson, E. B. and M. M. Hilferty (1931). The distribution of chi-square. Proceedings of the National Academy of Sciences of the United States of America 17(12), 684–688.
- Wirtschaftskammer Österreich (2017). Wirtschaftsdaten auf Bezirksebene. <https://www.wko.at/service/zahlen-daten-fakten/wirtschaftsdaten-bezirksebene.html>. [accessed: 07.02.2018].
- Withers, C. S. and S. Nadarajah (2007). Linear regression with extreme value residuals. Communications in Statistics - Simulation and Computation 37(1), 73–91.
- Wolter, K. (2007). Introduction to Variance Estimation. New York: Springer-Verlag.
- Wooldridge, J. (2000). Introductory Econometrics: A Modern Approach. Andover: Cengage Learning EMEA.
- World Bank Institute (2005). Introduction to poverty analysis. Technical report, The World Bank.

- Yale, C. and A. B. Forsythe (1976). Winsorized regression. Technometrics 18(3), 291–300.
- Yang, S. (1985). A smooth nonparametric estimator of a quantile function. Journal of the American Statistical Association 80(392), 1004–1011.
- Yang, Z. (2006). A modified family of power transformations. Economics Letters 92(1), 14–19.
- Yang, Z. and T. Abeyasinghe (2003). A score test for Box-Cox functional form. Economics Letters 79(1), 107–115.
- Yang, Z. and A. K. Tsui (2004). Analytically calibrated Box-Cox percentile limits for duration and event-time models. Insurance: Mathematics and Economics 35(3), 649–677.
- Yeo, I.-K. and R. A. Johnson (2000). A new family of power transformations to improve normality or symmetry. Biometrika 87(4), 954–959.
- Yoshimori, M. and P. Lahiri (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. Journal of Multivariate Analysis 124, 281–294.
- Yoshizawa, C., P. Sen, and E. Davis (1985). Asymptotic equivalence of the Harrel-Davis median estimator and the sample median. Communications in Statistics - Theory and Methods 14(9), 2129–2136.
- You, Y. and B. Chapman (2006). Small area estimation using area level models and estimated sampling variances. Survey Methodology 32(1), 97–103.
- Zar, J. H. (1999). Biostatistical Analysis. Upper Saddle River: Prentice Hall.
- Zarembka, P. (1974a). Frontiers in Econometrics. New York: Academic Press.
- Zarembka, P. (1974b). Transformation of variables in econometrics. In J. Eatwell, M. Milgate, and P. Newman (Eds.), Econometrics, pp. 257–260. London: Palgrave Macmillan.
- Zeckhauser, R. and M. Thompson (1970). Linear regression with non-normal error terms. The Review of Economics and Statistics 52(3), 280–286.
- Zeileis, A. (2014). ineq: Measuring Inequality, Concentration, and Poverty. R package version 0.2-13.
- Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics 57(3), 795–802.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B 67(2), 301–320.
- Zwet, W. R. (1964). Convex Transformations of Random Variables. Amsterdam: Mathematisch Centrum.

Summaries

Abstracts in English

Abstract: The Fay-Herriot model for multiply imputed data with an application to regional wealth estimation in Germany

The increasing inequality of private income and wealth requires the redistribution of financial resources. Thus, several financial support schemes allocate budget across countries or regions. This work shows how to estimate private wealth at low regional levels by means of a modified Fay-Herriot approach that deals with a) unit and item non-response, especially with used multiple imputation, b) the skewness of the wealth distribution, and c) inconsistencies of the regional estimates with the national direct estimate. One compelling example for financial redistribution is the promoted catching-up process of East Germany after the German reunification. This work shows that 25 years after the reunification differences are more diverse than just between the East and the West by estimating private wealth at two regional levels in Germany. The analysis is based on the Household Finance and Consumption Survey (HFCS) that the European Central Bank launched for all euro area countries in 2010. Although the application in this paper focuses particularly on Germany, the approach proposed is applicable to the other countries participating in the HFCS as well as to other surveys that make use of multiple imputation.

Keywords: small area estimation, non-response, multiple imputation, HFCS

Abstract: The fayherriot command for estimating small-area indicators

The command `fayherriot` implements the Fay-Herriot model (Fay and Herriot, 1979), a small-area estimation technique (Rao and Molina, 2015), in Stata. The Fay-Herriot model (Fay and Herriot, 1979) improves the precision of area-level direct estimates using area-level covariates. It belongs to the class of linear mixed models with normally distributed error terms. The `fayherriot` command encompasses options that a) produce out-of-sample predictions, b) adjust non-positive random effects variance estimates, and c) deal with the violation of model assumptions.

Keywords: disaggregated indicators, small area estimation, (log-transformed) Fay-Herriot model, empirical best linear unbiased predictor (EBLUP)

Abstract: Estimation of sample quantiles: Challenges and issues in the context of income and wealth distributions

Means, quantiles and extreme values are common statistics for the description of distributions. However, estimating sample quantiles with the default definition in different software programs leads to unequal results. This is due to the fact that software programs use different quantile definitions. Since most practitioners are not aware of this fact and use different quantile definitions interchangeably, this work compares the default definitions in the software programs SPSS, R, SASTM software, and Stata and additional quantile definitions that are suggested by the literature. The work especially focuses on how the quantile estimators perform in the context of describing the distribution of income and wealth. Furthermore, the possibilities of considering sampling weights in the quantile estimation and methods for producing variance estimates using the above-mentioned software are discussed.

Keywords: quantile definitions, software comparison, weighted quantile estimator, weighted Harrell-Davis estimator

Abstract: The R package emdi for estimating and mapping regionally disaggregated indicators

The R package **emdi** enables the estimation of regionally disaggregated indicators using small area estimation methods and includes tools for processing, assessing, and presenting the results. The mean of the target variable, the quantiles of its distribution, the headcount ratio, the poverty gap, the Gini coefficient, the quintile share ratio, and customized indicators are estimated using direct and model-based estimation with the empirical best predictor (Molina and Rao, 2010). The user is assisted by automatic estimation of data-driven transformation parameters. Parametric and semi-parametric, wild bootstrap for mean squared error estimation are implemented with the latter offering protection against possible misspecification of the error distribution. Tools for (a) customized parallel computing, (b) model diagnostic analyses, (c) creating high quality maps and (d) exporting the results to ExcelTM and OpenDocument Spreadsheets are included. The functionality of the package is illustrated with example data sets for estimating the Gini coefficient and median income for districts in Austria.

Keywords: official statistics, survey statistics, parallel computing, small area estimation, visualization

Abstract: A guideline of transformations in linear and linear mixed regression models

Representing a relationship between a response variable and a set of covariates is an essential part of the statistical analysis. The linear regression model offers a parsimonious solution to this issue, and hence it is extensively used in nearly all science disciplines. In recent years the linear mixed regression model has become common place in the statistical analysis. Numerous assumptions are usually made whenever these models are employed in scientific research. If one or several of these assumptions are not met, the application of transformations can be useful. This work provides an extensive overview of different transformations and estimation

methods of transformation parameters in the context of linear and linear mixed regression models. The main contribution is the development of a guideline that leads the practitioner working with data that does not meet model assumptions by using transformations.

Keywords: model assumptions, normality, transformation parameters, hierarchical models, multilevel analysis

Abstract: The R package `trafo` for transforming linear regression models

The linear regression model has been widely used for descriptive, predictive, and inferential purposes. This model relies on a set of assumptions, which are not always fulfilled when working with empirical data. In this case, one solution could be the use of more complex regression methods that do not strictly rely in the same assumptions. However, in order to improve the validity of model assumptions, transformations are a simpler approach and enable the user to keep using the well-known linear regression model. But how can a user find a suitable transformation? The R package `trafo` offers a simple user-friendly framework for selecting a suitable transformation depending on the user needs. The collection of selected transformations and estimation methods in the package `trafo` complement and enlarge the methods that are existing in R so far.

Keywords: power transformations, optimal parameter, model assumptions, normality

Kurzzusammenfassungen auf Deutsch

Zusammenfassung: Das Fay-Herriot Modell für mehrfach imputierte Daten angewendet auf die Schätzung von regionalem Vermögen in Deutschland

Die ansteigende ungleiche Verteilung von privatem Einkommen und Vermögen erfordert die Umverteilung von finanziellen Ressourcen. Daher wird im Zuge von Plänen zur finanziellen Unterstützung zwischen Ländern und Regionen Budget verteilt. Ein bekanntes Beispiel in diesem Kontext ist der geförderte Aufholprozess Ostdeutschlands nach der Wiedervereinigung. Allerdings ist 25 Jahre nach der Wiedervereinigung fraglich, ob Unterschiede wirklich nur zwischen Ost und West bestehen oder ob die Betrachtung von kleineren Regionen nicht ein genaueres Bild zum Vorschein bringt. Um eine Datengrundlage für die Schätzung von Privatvermögen zu haben, erhebt die Europäische Zentralbank seit dem Jahr 2010 die Household Finance and Consumption Survey (HFCS) in allen Ländern der Eurozone. Diese Arbeit stellt vor, wie Schätzer für deutsche Regionen (Bundesländer und Raumordnungsregionen) basierend auf der HFCS mit Hilfe eines modifizierten Fay-Herriot Modells berechnet werden können. Das vorgestellte Verfahren berücksichtigt a) die Schiefe der Vermögensverteilung, b) Unit und Item Non-Response, vor allem die angewandte Multiple Imputation, und c) Inkonsistenzen zwischen den regionalen Schätzern und dem direkten nationalen Schätzer. Obwohl die Arbeit sich auf Deutschland konzentriert, ist die vorgeschlagene Methode auch für die anderen Länder, in denen die HFCS durchgeführt wird, anwendbar, ebenso wie für Befragungen, die eine ähnliche Datenstruktur aufweisen.

Schlüsselwörter: Small-Area-Methoden, Non-Response, Multiple Imputation, HFCS

Zusammenfassung: Das `fayherriot` Kommando zur Schätzung von kleinräumigen Indikatoren

Das Kommando `fayherriot` ermöglicht die Anwendung des Fay-Herriot Modells (Fay and Herriot, 1979), das zu den Small-Area-Methoden gehört, in Stata. Das Fay-Herriot Modell (Fay and Herriot, 1979) erhöht die Präzision von aggregierten direkten Schätzern durch die Nutzung von aggregierten Kovariaten. Es gehört zur Klasse von linear gemischten Modellen mit normalverteilten Fehlertermen. Das `fayherriot` Kommando umfasst Optionen, die es ermöglichen a) Out-of-Sample Prädiktionen zu erhalten, b) negative Varianzschätzungen des zufälligen Effekts zu vermeiden, und c) Verletzungen von Modellannahmen entgegen zu wirken.

Schlüsselwörter: Disaggregierte Indikatoren, Small-Area-Methoden, (logarithmiertes) Fay-Herriot Modell, empirisch bester unverzerrter Prädiktor unter den linearen Schätzern (EBLUP)

Zusammenfassung: Die Schätzung von Quantilen: Herausforderungen und Probleme im Kontext von Einkommens- und Vermögensverteilungen

Mittelwerte, Quantile und Extremwerte sind übliche Statistiken, die zur Beschreibung von Verteilungen genutzt werden. Allerdings sind die Ergebnisse für Quantile, die mit verschiedener Software berechnet werden, nicht zwingend gleich. Dies ist darauf zurückzuführen, dass Quantilsdefinitionen, die in verschiedenen Software-Programmen genutzt werden, nicht einheitlich sind. Da diese unterschiedlichen Definitionen vielen Anwendern nicht bewusst sind, vergleicht diese Arbeit die unterschiedlichen Quantilsdefinitionen der Software-Programme SPSS, R, SASTM Software und Stata. Außerdem werden Quantilsdefinitionen betrachtet, die in der Literatur evaluiert und empfohlen wurden. Diese Arbeit betrachtet besonders die Güte der unterschiedlichen Quantilsdefinitionen für die Beschreibung von Einkommens- und Vermögensverteilungen. Außerdem werden Möglichkeiten zur Berücksichtigung von Survey-Gewichten bei der Quantilsschätzung, sowie zur Varianzschätzung in den genannten Software-Programmen diskutiert.

Schlüsselwörter: Quantilsdefinitionen, Vergleich von Software, gewichtete Quantilsschätzer, gewichteter Harrell-Davis Schätzer

Zusammenfassung: Das R Paket `emdi` für die Schätzung und die Erstellung von Karten für regional disaggregierte Indikatoren

Das R Paket `emdi` ermöglicht die Schätzung von regional disaggregierten Indikatoren mittels Small-Area-Methoden und enthält Funktionen für die Erstellung, die Analyse und die Präsentation von Schätzergebnissen. Der Mittelwert, die Quantile der Verteilung, die Armutsquote, die Armutslücke, der Gini-Koeffizient und das Quintilsverhältnis, sowie individuell definierte Indikatoren können mit direkter Schätzung oder modellbasierten Verfahren, mit dem Empirical Best Predictor (Molina and Rao, 2010), geschätzt werden. Der Anwender wird dabei durch die automatische Schätzung von Transformationsparametern für datengetriebene Transformationen unterstützt. Ein parametrischer und ein semi-parametrischer wild Bootstrap für die Schätzung des mittleren quadratischen Fehlers sind implementiert, wobei der zweite zusätzlich gegen die mögliche Misspezifikation der Fehlerverteilung schützt. Das Paket ermöglicht (a)

parallele Berechnungen, (b) die Analyse von Modellannahmen, (c) die Erstellung von Karten, (d) den Export von Ergebnissen zu Excel™ und zu OpenDocument Spreadsheets. Die Funktionalität des Pakets wird mit der Schätzung des Gini-Koeffizienten und des Medians für österreichische Bezirke basierend auf Beispieldatensätzen illustriert.

Schlüsselwörter: Amtliche Statistik, Survey-Statistik, parallele Berechnungen, Small-Area-Methoden, Visualisierung

Zusammenfassung: Ein Leitfaden für die Nutzung von Transformationen in linearen und linear gemischten Modellen

Ein großer Bestandteil statistischer Analysen besteht darin, den Zusammenhang zwischen einer abhängigen und mehreren erklärenden Variablen zu beschreiben. Da das lineare Regressionsmodell eine einfache Lösung für die Beschreibung dieses Zusammenhangs ist, wird es in vielen Wissenschaften angewandt. Seit einiger Zeit werden auch immer häufiger linear gemischte Regressionsmodell genutzt. Beide Modelltypen basieren auf Annahmen, die bei der Anwendung überprüft werden und erfüllt sein sollten. Wenn eine oder mehrere dieser Annahmen nicht erfüllt sind, können Transformationen helfen weiterhin die Modellklasse der linearen Modelle zu nutzen. Dafür bietet diese Arbeit einen umfassenden Überblick über verschiedene Transformationen und Schätzmethoden für die Schätzung eines optimalen Transformationsparameters basierend auf den zugrunde liegenden Daten im Kontext von linearen und linear gemischten Modellen. Der größte Beitrag der Arbeit liegt darin, dem Anwender Leitlinien an die Hand zu geben, wie man Transformationen nutzen kann, um die Modellannahmen des linearen Modells zu erfüllen, und was dabei beachtet werden muss.

Stichworte: Modellannahmen, Normalität, Transformationsparameter, hierarchische lineare Modelle, Mehrebenenanalyse

Zusammenfassung: Das R Paket `trafo` für die Transformation von linearen Modellen

Das lineare Regressionsmodell ist eine häufig genutzte statistische Methode, um Zusammenhänge zu beschreiben und Vorhersagen durchzuführen. Allerdings beruht das Modell auf Annahmen, die in der Anwendung nicht immer erfüllt sind. In diesen Fällen könnten zum einen komplexere Methoden genutzt werden, die nicht auf den gleichen Annahmen beruhen. Zum anderen können Transformationen helfen, um die Gültigkeit der Annahmen zu verbessern. Um eine passende Transformation zu finden, bietet das R Paket `trafo` einen anwenderfreundlichen Rahmen. Die Auswahl an Transformationen und Schätzmethoden für den Transformationsparameter in diesem Paket ergänzen die bisher angebotenen Methoden in R.

Stichworte: Power Transformationen, optimaler Transformationsparameter, Modellannahmen, Normalität

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Berlin, January 25, 2019

Ann-Kristin Kreutzmann
January 25, 2019