Thesis submitted in fulfillment of the requirements for the degree

Dr. rer. pol.

on the topic

Combining Data Sources: A Path to Improved Understanding and Prediction

to the

Chair of Applied Statistics

School of Business and Economics

Freie Universität Berlin

submitted by

Sören Pannier

born in Hamburg

Berlin, 2021

Sören Pannier, *Combining Data Sources: A Path to Improved Understanding and Prediction*,
February 2021

Supervisors:
Prof. Dr. Timo Schmid (Freie Universität Berlin)
Prof. Nikos Tzavidis, Ph.D. (University of Southampton)
Prof. Dr. Ulrich Rendtel (Freie Universität Berlin)

Location:
Berlin

Date of defense: July 22, 2021

Acknowledgements

I want to express my deepest gratitude to my supervisor, Prof. Dr. Timo Schmid (Freie Universität Berlin, Germany), whose guidance and support were invaluable for my Thesis.

I would also like to convey my appreciation to Prof. Nikos Tzavidis Ph.D. (Southampton University, England) and Prof. Dr. Ulrich Rendtel (Freie Universität Berlin, Germany) for all their input, collaborations, and long and fruitful discussions.

Furthermore, I thank my friends and colleagues at the Chair of Statistics and *fu:stat* for their insightful suggestions and constructive advice, as well as for their cheerful distractions.

Last but not least, I want to thank my partner and my family for their unconditional love and support.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "Combining Data Sources: A Path to Improved Understanding and Prediction."

1. Pannier, S., Rendtel, U. and Gerks, H. (2020), **Die Prognose von Studienerfolg und Studienabbruch auf Basis von Umfrage- und administrativen Prüfungsdaten.** AStA Wirtsch Sozialstat Arc, 14: 225–266. doi.org/10.1007/s11943-020-00278-5

2. Rojas-Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N. (2020), **Data-driven transformations in small area estimation.** J. R. Stat. Soc. A, 183: 121-148. doi.org/10.1111/rssa.12488

3. Harmening, S., Kreutzmann, A.-K., Pannier, S., Salvati, N. and Schmid, T. (2020), **A Framework for Producing Small Area Estimates Based on Area-Level Models in R.** *Working paper*, submitted to Journal of Statistical Software

4. Pannier, S (2020) **ammlogit: Estimation and Prediction Using an Aggregated Mixed Multinomial Logit Model.** *Working paper*, to be submitted

5. Hensel, S., Pannier, S., Schmid, T., and Tzavidis, N. (2020), **Asymptotic distribution of regression quantiles in a mixed effects model.** *Working paper*, to be submitted

Contents

Introduction

Over the last few years, data has often been described as the oil of the 21st century, e.g., by Bhageshpur (2019). Just as access to oil dominated power and development in the last century, this claim implies that personal data is not only assumed to be similarly valuable, but also equally as influential in politics and society as oil once was. However, oil sources mainly diverge in their accessibility, quality, quantity, and cost of exploitation but, once extracted and refined theses sources may lead to roughly similar products. Data also differs in these four categories but, additionally, data sources typically lead to very specific insights. One single data source is often neither sufficient to answer important and complex scientific (or economic) questions nor to make any predictions with fine granularity and high precision. In such cases, combining data from different sources that provide additional aspects to the problem at hand is one promising approach to achieve these aims.

In this dissertation, combining data sources is conducted for two purposes. Part I of this work focuses on combining data to achieve additional understanding. In the paper presented in Part I, the authors analyze reasons why students drop out of undergraduate courses in economics and business administration. From a university perspective, administrative data is readily available, e.g., which modules are completed in which semester, how many educational credit points are achieved by each student in each semester. Socioeconomic data at individual level, however, is usually unavailable to university administrations. In order to overcome this hurdle, the authors proposed and executed a novel prospective study design. A survey was conducted on students starting the second semester and the data was combined with administrative longitudinal data. Hence, the authors were able to analyze individual studying behavior conditioned on a large pool of socio-demographic variables. Among other results, the authors were able to show that college admission grades have a negligibly small impact on the achievements made on Bachelor degree courses. This finding stands in strong contrast to college admission policies in Germany that strongly focus on high-school grades for college admission.

The second purpose of data-combination, as discussed in this work, is the combination of data-sources to improve the precision of predictions. Part II consists of three papers from the field of small area estimation (SAE). In SAE problems, there is some survey data typically available that contains the (VOI). However, an indicator of interest needs to be estimated on some subgroup level by a function of the VOI. Such levels usually consist of a geographic

region but are not limited to this. In the context of SAE, these subgroups are called areas. With an increasing number of areas, the quantity of observations available per area decreases, often leading to areas with very few or even without any observations (out-of-sample areas). In such situations, a prediction with reasonable reliability becomes impossible when only relying on the available survey data. One possibility to overcome this burden is to couple the survey data with additional data, such as administrative or census data. Frequently, these data sources do not contain the VOI, thus rendering a direct estimation of the indicator impossible. However, if similar covariates are available in the survey and census, a feasible approach is to assume a model for the VOI that holds, both in the survey and census, to estimate the model on the survey data and then to combine the model estimates with the more numerous census data that enables prediction. This general method is often referred to as "borrowing strength." Such model-based approaches are roughly divided into two classes, depending on the data availability and resulting requirements on the models. First, if the data is available for each individual of interest, e.g., a citizen or household, unit-levels are used. For area-level models, on the other hand, data is only available in aggregated form at area level.

Many common methods in SAE assume normally distributed residuals, not only for the model estimation, but more crucially when using the census data for prediction and the estimation of precision. Therefore, deviations from normality have severe consequences and lead to less precise point estimates and misleading shrunk error estimates. The paper presented in Chapter 2 proposes the use of data-driven transformations, resulting in smaller deviations from normality and thus improved point and precision estimates. Chapter 3 presents the R-package **emdi** that implements not only the latter methodology, but also allows for the usage of various area-level Fay-Herriot models. **emdi** focuses on user-friendliness and provides many useful functionalities to support the user through every step, from model estimation through the analysis of model assumptions to visualize the results and enable their exportation. However, when deviations from normality are too severe, different model types are more suitable. Chapter 4 presents the R-package **ammlogit**, which allows the user to work with a multinomial VOI. The corresponding paper introduces a new methodology for prediction and a revised bootstrap mean squared error estimation. Like **emdi**, **ammlogit** is designed to be user-friendly and narrow the gap between research and practitioners. Still, the method used in **ammlogit** also uses distributional assumptions, as the counts are assumed to be conditionally multinomially distributed and the area-level error terms are assumed to be identically normally distributed.

A model class without distributional assumptions are quantile-type regression models. M-Quantile models have been used in SAE since Chambers and Tzavidis (2006). However, the related mixed quantile regression models have not yet been consequently applied to small area problems, even though they are a naturally robust alternative to the widely used linear mixed regression models that dominate research and application. In parts, this may be due to remaining uncertainties about their asymptotic properties. Therefore, in the paper presented in Part III, the asymptotic normality of the corresponding maximum likelihood estimates is proven and a plugin-variance estimator is derived.

Part I

Combining Data for Understanding

Chapter 1

Studienverläufe

1.1 Einleitung

Studienabbrüche und deren Ursachen waren schon immer im Fokus der Evaluierung von Hochschulen, vgl. den Überblick von Neugebauer u. a. (2019). So hat das Deutsche Zentrum für Hochschul- und Wissenschaftsforschung (DZHW) hierzu mehrfach breit angelegte Befragungen unter Studienabbrechern und Studienabsolventen vorgelegt, vgl. Heublein u. a. (2017) Hierbei handelt es sich um retrospektive Befragungen, in denen die Befragten über ihr früheres Studium berichten. Ihre Auswahl erfolgte über zwei Stichproben: Eine Stichprobe aus dem Kreis der Exmatrikulierten, die als Grund „Studienabbruch“ oder „endgültig nicht bestandene Prüfung“ angaben, sowie eine Stichprobe aus dem Kreis der erfolgreichen Studienabsolventen. Die Stichprobenmitglieder wurden postalisch ein halbes Jahr nach Verlassen der Hochschule kontaktiert. Die Autoren geben eine Rücklaufquote von 23% an.

Die „Konstanzer Studiensurveys“ verwenden ein querschnitts-orientiertes Befragungsdesign (Georg, 2008). Hier wird während des Studiums nach der Absicht der Studierenden gefragt, ob sie eventuell an einen Studienabbruch denken. Die Befragung erfolgt wiederum postalisch durch Versendung eines Fragebogens mit einer Rücklaufquote von 35%.[1] Diese Befragung wird in regelmäßigen zeitlichen Intervallen wiederholt.

Isphording und Wozny (2018) benutzen dagegen einen Panel-Ansatz. Sie benutzen die 2010/11 Kohorte der Studienanfänger im Nationalen Bildungspanel (NEPS). Analysiert wird ein Zeitraum von 10 Semestern. Die Befragung streut breit über Fächer und Hochschulen.

In diesem Aufsatz präsentieren wir zu diesem Thema ein neues Erhebungsdesign. Es ist ein prospektives Erhebungsdesign, das nahezu ohne Verluste durch Nonresponse auskommt. Ein prospektiver Ansatz ist zwar schon in früheren Arbeiten zum Studienerfolg benutzt worden, vgl. die Arbeiten von Bean (1982), Gold (1988), Brandstätter u. a. (2006) sowie Fleischer u. a.

[1] An der HU Berlin wird nach demselben Konzept eine schriftliche Befragung durchgeführt. Dort lag die Rücklaufquote nie über 10%. In 2015 lag sie bei 2%.

(2019). Allerdings weisen die hierbei benutzten Erhebungsdesigns gewisse Einschränkungen auf. So wird der Studienverlauf nur auf Basis von Studierenden ermittelt, die vorher an einer schriftlichen Befragung teilgenommen haben, wie beispielsweise bei Bean (1982). Hier ist der zeitliche Abstand zwischen der Befragung und der Ermittlung der Studienleistung mit einem Jahr relativ kurz. Bei Brandstätter u. a. (2006) werden die Studierenden vor Aufnahme des Studiums und zu einem späteren Zeitpunkt während des Studiums befragt.

Prospektive Studien ohne Verluste durch Survey-Nonresponse wie etwa Berens u. a. (2019), Danilowicz-Gösele u. a. (2017), Schneider u. a. (2019) oder Fleischer u. a. (2019) können in Deutschland seit der breiten Umstellung der Prüfungsverwaltungen auf eine moderne EDV-Infrastruktur auf Basis der administrativen Daten analysiert werden. Allerdings sind hier die datenschutzrechtlichen Hürden hoch und es fehlen Angaben zur Studienmotivation, zum sozialen Hintergrund und zur ökonomischen Lage der Studierenden. Häufig – jedoch nicht immer – können Angaben aus dem Zulassungsverfahren, wie etwa die Abiturnote, das Bundesland der Hochschulberechtigung neben dem Geschlecht und dem Alter für die Analysen genutzt werden. Bei Berens u. a. (2019, Table 1) werden die Merkmale aus dem Hochschulstatistik-Gesetzes dazu genutzt, um den sozialen Hintergrund der Studierenden über die Art der Krankenversicherung und den lokalen Kaufkraftindex zu erschließen.

In diesem Aufsatz berichten wir über ein Pilotprojekt, das im Sommersemester 2016 am Fachbereich Wirtschaftswissenschaft der Freien Universität Berlin gestartet wurde. Es war das Ziel des Projekts „Studienverläufe“, die Studierenden direkt im Hörsaal zu erreichen, sie über einen Fragebogen zu befragen sowie sie um ihre Einwilligung zu bitten, die Antworten mit Ihren Prüfungsdaten verknüpfen zu dürfen. Diese Einwilligung wurde in 95% aller Fälle gegeben. Eine hohe Responserate bei den Fragebögen wurde erzielt, weil die Studierenden bei Start einer Pflichtveranstaltung („Statistik für Wirtschaftswissenschaftler“) direkt im Hörsaal angesprochen wurden. Zu diesem Zeitpunkt befanden sich die Studierenden am Anfang des zweiten Fachsemesters. Sie hatten also schon Erfahrungen mit ihrem Studium (Bachelor Betriebswirtschaftslehre (BA BWL) und Bachelor Volkswirtschaftslehre (BA VWL) gemacht und konnten sich über mögliche Absichten zur Fortsetzung der Studiums äußern beziehungsweise über eine Absicht, den Studiengang zu verlassen.

Dieser verlaufsorientierte Ansatz bietet einige Vorteile gegenüber dem retrospektiven Ansatz von Heublein u. a. (2017) sowie dem querschnittsorientierten Ansatz von Georg (2008). Zum einen liegt nach einer gewissen Wartezeit der vollständige Prüfungsverlauf bis zur Exmatrikulation vor. Man kann also überprüfen, ob Studienabbrecher an vielen Prüfungen erfolglos teilgenommen haben oder ob sie sich eher aus dem Prüfungsbetrieb zurückgezogen haben. In dieser reichhaltigen Fülle von objektiven Leistungsmessungen liegt auch der Vorteil gegenüber der Survey-Messung im Rahmen eines Panel-Ansatzes von Isphording und Wozny (2018). Hinzu kommen bei Panel-Surveys die Ausfälle im Verlauf des Panels, die sogenannte Panel-Mortablität. Hier belegt beispielsweise Schimpl-Neimanns (2008), dass Bildungsverläufe gerade in Wechselsituationen nicht ignorierbaren Ausfallmustern unterliegen und Wechsel un-

terschätzt werden. Basic und Rendtel (2007) belegen einen selektiven Einfluss von Umzügen auf Analysen des Erwerbsverhaltens mit dem Mikrozensus. Der hier präsentierte Ansatz bleibt von diesem Problem unberührt. Der Ansatz über die Hörsaalbefragung in einer Pflichtveranstaltung ermöglicht eine sehr hohe Teilnahmerate und minimiert vor allem die Verluste durch fehlende schriftliche Einwilligungserklärungen zur Verknüpfung mit den Prüfungsdaten. Der Vorteil, Studierende direkt in Veranstaltungen anzusprechen, wurde auch von Himmler u. a. (2019) genutzt. Hierbei wurde eine Einführungsveranstaltung[2] genutzt, um die Studierenden auf ein Commitment hinsichtlich des Studienabschlusses anzusprechen.

Die individuelle Zufriedenheit mit dem Studium spielt neben der sozialen und akademischen Integration (Tinto (1975)) der Studierenden eine zentrale Rolle in der Literatur zu Studienerfolg und Studienabbruch, vgl. Gold (1988), Wiers-Jenssen u. a. (2002), Schiefele und Jacob-Ebbinghaus (2006), Brandstätter u. a. (2006) und Multrus u. a. (2017). Auch aus diesem Grund erscheint eine Messung dieser zentralen Größe zu Beginn des zweiten Semesters[3] angemessen zu sein, da die Studierenden nun ein vollständiges Bild über ihr erstes Semester und den dabei erzielten Studienerfolg[4] haben.

Während sich viele Analysen auf die Kausalität und die Rolle der Einflussfaktoren auf den Studienabbruch richten, ist die Identifikation potentieller Studienabbrecher eine hiervon verschiedene Aufgabe. Eine frühe, zuverlässige Prognose eines Studienabbruchs ist ein wertvolles Hilfsmittel für Mentorenprogramme. Fleischer u. a. (2019) haben in diesem Zusammenhang die Äußerung eines möglichen Studienabbruchs als Prognoseinstrument untersucht. Sie kommen für die naturwissenschaftlichen Fächer zu dem Ergebnis, dass die Studienergebnisse am Ende des ersten Semesters die Prognosen erheblich verbessern. Unter einem anderen Aspekt untersuchen Berens u. a. (2019) administrative Prüfungsdaten auf ihre Prognosefähigkeit mit Hilfe von Machine Learning Methoden und kontrastieren sie mit dem Standard-Ansatz über ein Logit-Modell. Mit unserem Erhebungsdesign stehen uns zusätzlich soziodemographische Hintergrundmerkmale sowie die Zufriedenheit mit dem Studium für die Prognose zur Verfügung.

Schließlich bietet die Kenntnis der individuellen Prüfungsprofile in Verbindung mit den Hintergrundvariablen zahlreiche weitere Analysemöglichkeiten: Etwa, wie schnell Verzögerungen gegenüber dem Studienplan wieder aufgeholt werden oder bis zu welchem Grad den Empfehlungen des Studienplans überhaupt gefolgt wird. Eine erste Beschreibung der Möglichkeiten, administrative Prüfungsdaten zu nutzen findet man bei Hahm und Storck (2018).

Die Verallgemeinerbarkeit der Befunde ist bei diesem Erhebungsdesign jedoch dadurch eingeschränkt, dass nur Befunde aus dem Prüfungsregister einer Universität für die Analyse zu Verfügung stehen. Im der hier vorgestellten Analyse war es jedoch möglich, einige Analysen

[2] In einer ähnlichen Umfrage zum Erfolg in 5 Masterstudiengänge am Fachbereich haben wir die Studierenden ebenfalls auf der Einführungsveranstaltung des jeweiligen Studiengangs angesprochen.

[3] Allerdings hat dies die Konsequenz, dass Studierende, die ihr Studium schon vor dem zweiten Semester abbrechen, nicht berücksichtigt werden.

[4] Die Prüfungsergebnisse liegen in der Regel erst unmittelbar vor Beginn des neuen Semesters vor.

an einer zweiten Berliner Universität, der Humboldt Universität zu Berlin[5] im selben Studienfach für dieselben Studienkohorten zu replizieren, so dass Rückschlüsse über die einzelnen Universitäten hinaus möglich sind..

Der Artikel ist wie folgt gegliedert: Im folgenden zweiten Abschnitt beschreiben wir die Erzeugung des Datensatzes; insbesondere die datenschutzrechtlichen Aspekte der Zusammenführung der Prüfungsdaten mit den Umfragedaten. Der dritte Abschnitt präsentiert einige beschreibende Darstellungen der Studienverläufe. Schließlich untersuchen wir in Abschnitt 4 in wie weit sich zeitliche Defizite gegenüber dem Studienplan im Laufe des Studiums verstärken oder verringern. Im fünften Abschnitt beschäftigen wir uns mit der Studieneingangsphase, konkret mit den erworbenen Studienpunkten während des ersten Studienjahrs und ihrer Prognosekraft auf die Studienwechselneigung (Georg, 2008) und einen späteren Abbruch des Studiums. Im 6. Abschnitt prüfen wir den Einfluss der Abiturnote auf die universitäre Performance der Studierenden, also den Abschluss des Studiums und die hierbei erzielte Note im Bachelor. Im Resümee ziehen wir Folgerungen für die Betreuung der Studierenden sowie für die Zulassung zum Studium.

1.2 Die Erzeugung des Datensatzes

Für die Realisierung des Projekts waren einige Voraussetzungen notwendig. Auf der Seite der zentralen Prüfungsverwaltung musste es möglich sein, individuelle Studienverläufe aus einer Datenbank für bestimmte Studiengangskohorten abzurufen. Diese Voraussetzung ist beispielsweise im lokalen Prüfungsbüro des Fachbereichs Wirtschaftswissenschaft nicht gegeben, da dort nur das aktuelle Prüfungskonto einzelner Studierender abgerufen werden kann[6] Weiterhin gestatten Datenschutzvorbehalte[7] nur die Verwendung einer pseudonymisierten Matrikelnummer, die zwar über die Zeit eindeutig und fix ist, allerdings keine Rückschlüsse auf die wahre Matrikelnummer zulässt.

Weiterhin muss die Zusammenführung der Fragebogeninhalte und der Studienverlaufsinformation in einer datenschutzrechtlich abgesicherten Weise erfolgen. Die Basis hierfür ist eine schriftliche Einwilligungserklärung der Studierenden, die über den Zweck der Untersuchung, die Verwendung der Daten sowie den Abschluss des Projekts aufklärt. Diese Einverständniserklärung enthielt die Fragebogennummer. Die Erklärung wurde im Anschluss der Hörsaalbefragung vom Fragebogen abgetrennt. Auf der Einverständniserklärung[8] konnten die Studierenden ihre wahre Matrikelnummer eintragen. Die separaten Einverständniserklärungen wurden dann der Datenschutzbeauftragten der FU als Treuhänderin übergeben. Diese leitete die echte Matrikelnummern an die Prüfungsverwaltung weiter, die als einzige Stelle den Umstiegs-

[5] Allerdings fehlen dort die Hintergrundmerkmale aus der Hörsaalbefragung.

[6] Die gesamte Prüfungsadministration der FU wird über SAP verwaltet. Das System gestattet daher den einzelnen Nutzern nur bestimmte, vorher festgelegte Auswertungen der eigentlichen Prüfungsdatenbank.

[7] Diese Vorbehalte gelten natürlich nur für Nutzer außerhalb der Prüfungsverwaltung. Umgekehrt ist es für die Mitarbeiter der Prüfungsverwaltung schwierig, an die Hintergrundgrundmerkmale der Studierenden zu kommen

[8] Die Einverständniserklärung ist im Anhang A dokumentiert. Da sie im Sommersemester 2016 eingesetzt wurde, fehlt dort ein Hinweis auf die später eingeführte Datenschutzgrundverordnung.

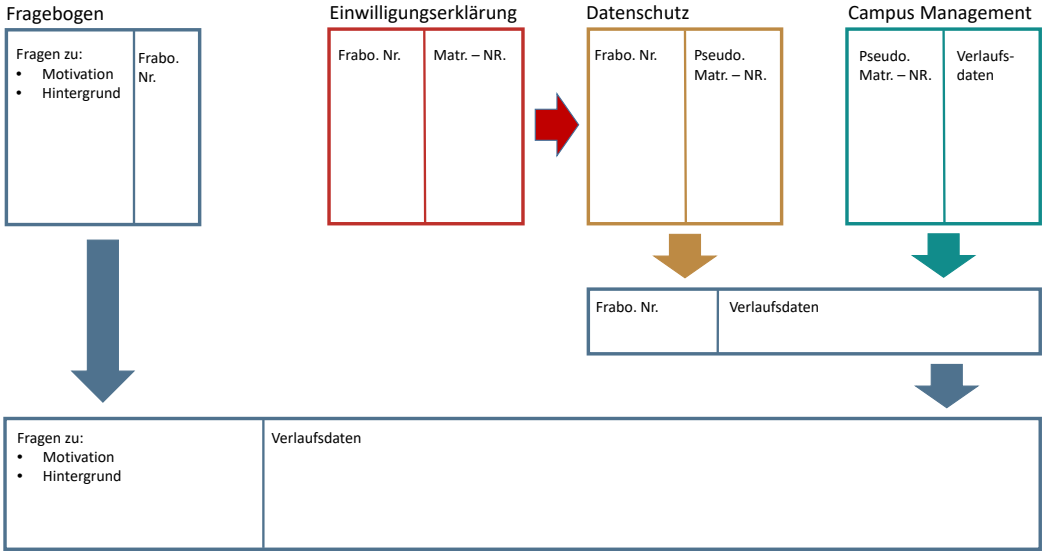


Abbildung 1.1: Die Zusammenführung der Umfragedaten mit den Prüfungsverläufen

schlüssel zu den pseudonymisierten Matrikelnummern besitzt. Von der Prüfungsverwaltung haben wir dann eine Liste erhalten, die jeder pseudonymisierten Matrikelnummer die Nummer des zugehörigen Fragebogens zuordnet. Anhand dieser Daten konnten dann die Studienverläufe und die Fragebögen zusammengeführt werden. Abbildung 1.1 stellt das Verfahren noch einmal im Zusammenhang dar.

Befragt wurden im Sommersemester 2016 insgesamt 322 Studierende, die überwiegend im Wintersemester 2015/16 ihr Studium aufgenommen hatten. Von diesen 322 Teilnehmern erhielten wir 305 (=0.95%) Einwilligungserklärungen zur Zusammenführung mit den Prüfungsverläufen. Da in der Vorlesung auch einige Teilnehmer aus anderen Kohorten anwesend waren, war die Anzahl der zusammengeführten Verläufe aus der Zielkohorte WS 2015/16 mit 233 etwas geringer. Insgesamt wurden 53% der Zielkohorte erreicht. Die Ausfälle erklären sich damit aus Verlusten, die vor dem zweiten Semester eingetreten sind, aus dem Nichtbesuch der Auftaktveranstaltung im Pflichtmodul Statistik für Wirtschaftswissenschaftler, aus Nonresponse beim Ausfüllen des Fragebogens, aus einer fehlenden Einwilligungserklärung sowie aus einer unleserlichen Angabe der Matrikelnummer, die ein Matching mit den Daten der Prüfungsverwaltung verhinderte.

Ein klassischer Ansatz zur Überprüfung der Repräsentativität der erhobenen Stichprobe vergleicht die Stichprobenverteilung mit bekannten Verteilungen der Population. Dies ist in diesem Fall die Verteilung der Studierenden am FB Wirtschaftswissenschaft nach Geschlecht und Fachrichtung.

Mit den Daten von Tabelle 1.1 ergeben sich keine signifikanten Unterschiede zu der Population am Fachbereich Wirtschaftswissenschaft. Reduziert man allerdings die Population auf das zweite Fachsemester, so zeigt sich eine signifikante Überrepräsentation der BWL-er mit 67%

	FB Wiwiss		2. Fachsem.	Stichprobe	
	Weibl.	Männl.		Weibl.	Männl.
BWL	519	454	255	127	82
VWL	182	358	188	36	66
Summe	1513		443	311	

Tabelle 1.1: Vergleich der Stichprobenverteilung mit der Verteilung nach Geschlecht und Studiengang am FB Wirtschaftswissenschaft

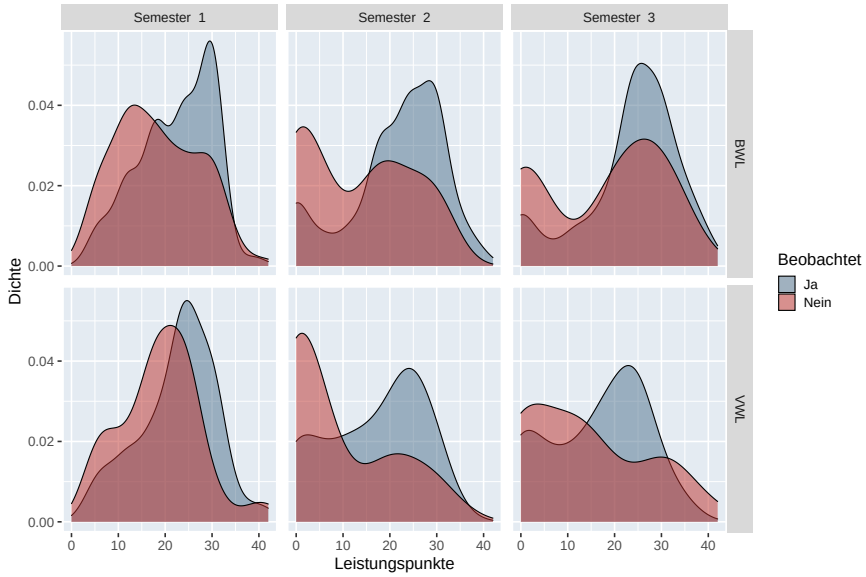


Abbildung 1.2: Vergleich der erworbenen Leistungspunkte (LP) zwischen Respondenten und Nonrespondenten in den ersten drei Fachsemestern (Oben BWL, unten VWL). Nonrespondenten mit $0 < LP < 42$ im ersten Fachsemester. Kerneldensity Schätzung mit automatischer Wahl des Glättungsparameters.

gegenüber 58% in der Population. Da aber für alle Analysen das Geschlecht und der Studiengang als Kontrollvariablen benutzt wurden, fällt diese Überrepräsentation der BWL-er nicht weiter ins Gewicht.

Der Zugang zu den administrativen Daten eröffnet jedoch weitere Möglichkeiten der Nonresponse-Kontrolle. So konnten wir die Verteilung der erreichten Leistungspunkte (LP) für die Respondenten mit denen der Nonrespondenten vergleichen. Um die Vergleichbarkeit zu erhöhen, wurden untypische Studierende unter den Nonrespondenten ausgeschlossen. Dies betrifft Studierende, die im ersten Fachsemester keine Leistungspunkte erworben habe. Hierbei handelt es sich meist um Studierende, die nach Semesterbeginn einen anderen Studienplatz im Nachrückverfahren erhalten haben. Weiterhin wurden Studierende ausgeschlossen, die mehr als 42 LP im ersten Semester erworben haben. Hierbei handelt es sich typischerweise um Studierende, die Leistungen aus anderen Studiengängen in den Bachelor eingebracht haben.

Abbildung 1.2 vergleicht die Kerndichteschätzer der Respondenten (Beobachtet=ja) und der Nonrespondenten (Beobachtet=nein). Hierbei zeigt sich schon im ersten Fachsemester (Semes-

ter=1), dass die Antwortenden etwas erfolgreicher im Studium waren als die Nonrespondenten. Diese Tendenz vergrößert sich in den folgenden beiden Semestern deutlich. Hierbei zeigt sich eine deutliche bimodale Struktur. Sowohl bei den Teilnehmern an der Befragung als auch bei den Nichtteilnehmern gibt es ein Cluster von Studierenden, die kaum noch Leistungspunkte erwerben, das sich deutlich von dem zweiten Cluster unterscheidet, wo Studierende Leistungspunkte im geforderten Bereich erwerben. Allerdings ist der Anteil des ersten Clusters bei den Nichtteilnehmern an der Befragung deutlich größer. Dies ist ein deutlicher Hinweis darauf, dass die engagierteren Studierenden auch befragungsbereiter sind. Ein derartiger "Erfolgsbias" wurde von Kühn u. a. (2018) für die empirische Bildungsforschung postuliert und ist in diesem Fall klar erkennbar.

Ohne eine Kontrolle, zum Beispiel über die erreichten Leistungpunkte, sind damit Aussagen über den Studienerfolg allein auf Basis der Stichprobe zu optimistisch. Auch aus diesem Grund haben wir uns bei der folgenden Ermittlung der Abbrecherquoten ganz auf eine Längsschnittauswertung der administrativen Daten verlassen. Bei der Ermittlung der Abbruchrisiken haben wir jedoch immer die erworbenen Leistungspunkte als Kontrollvariable benutzt.

Damit die Studienverläufe bis zum 8. Semester Eingang in die Analyse finden, muss bis zum Sommersemester 2019 auf die Prüfungs- und Rückmeldedaten gewartet werden. Diese zeitliche Differenz ist dem prospektiven Ansatz geschuldet. Allerdings wurden für frühere Kohorten Prüfungsverläufe bis zum 12. Fachsemester erstellt. Dies betrifft die Studienanfänger des Wintersemesters 2010/11, die im Sommersemester 2016 ihr 12. Fachsemester abgeschlossen haben. Allerdings fehlen für diese Kohorten die Hintergrundmerkmale.

Ungefähr 2/3 der Studierenden[9] war im Bachelor BWL eingeschrieben. Der Rest studierte im Bachelor VWL. Bei den folgenden Analysen werden diese beiden Studiengänge häufig separat analysiert, um Unterschiede im Studienerfolg, in der Studienmotivation und in den Erwartungen der Studenten darzustellen.

Um einige Ergebnisse auch über den Rahmen des Fachbereichs Wirtschaftswissenschaft der FU abzusichern, haben wir im Rahmen einer Kooperation mit der HU für die gleiche Kohorte Ergebnisse über Studienverläufe an der dortigen Wirtschaftswissenschaftlichen Fakultät ausgetauscht. Allerdings fehlen hier die Hintergrundvariablen aus der Umfrage. Nur hinsichtlich der erzielten Noten sind die Informationen aus der HU umfangreicher. Diese fehlen in unserer Untersuchung[10].

Insgesamt werden in dieser Analyse also verschiedene Datensätze analysiert, die in der folgenden Übersicht noch einmal zusammengestellt werden:

1. Datensatz 1: Kohorte WS2015/16 FU: Teilnehmer an Befragung im 2. Fachsemester, im

[9] 203 im BA-BWl und 119 im BA-VWL.

[10] Der Grund liegt in den unterschiedlichen Zugangsebenen: Während die Kooperationspartner aus der HU als Stabsstelle Qualitätsmanagement direkt auf die Prüfungsergebnisse zugreifen kann, hatten die Autoren an der FU nur einen eingeschränkten Zugriff über eine SAP-Schnittstelle.

SS2019 im 8. Fachsemester. Mit Hintergrundmerkmalen. Nur Studienpunkte und Studienstatus.

2. Datensatz 2: Kohorten WS2010/11 bis WS2013/14 FU: Seit 1. Fachsemester, Kohorten sind bei Abfassung des Manuskripts schon im 12. Fachsemester. Ohne Hintergrundmerkmale, Nur Studienpunkte und Studienstatus, Kein Zugang zu Noten. Darstellung von vollständigen individuellen Studienprofilen bis 12. Semester.

3. Datensatz 3: Kohorte WS2015/16 HU: Seit 1. Fachsemester, Ohne Hintergrundmerkmale, Mit Zulassungsmerkmalen, mit Studienpunkten und Studienstatus, mit detailliertem Prüfungsverhalten und Noten. Direkte Vergleiche mit FU Datensatz und Ergänzung zu Aussagen über Noten.

4. Datensatz 4: Kohorten WS2010/11 bis WS2013/14 HU: Seit 1. Fachsemester, Kohorten sind bei Abfassung des Manuskripts schon im 12. Fachsemester. Ohne Hintergrundmerkmale, Mit Zulassungsmerkmalen, mit Studienpunkten und Studienstatus, mit detailliertem Prüfungsverhalten und Noten. Einfluß von Zulassungsnoten auf Modul- und BA-Noten.

1.3 Deskriptive Darstellung von Studienverläufen

Eine Möglichkeit den Studienerfolg im Längsschnitt darzustellen, besteht darin, die jeweils erreichten Studienpunkte bis zum Studienabschluss bzw. bis zur Exmatrikulation dazustellen. Abbildung 1.3 zeigt die Summe der individuell erreichten Studienpunkte bis zum 12. Fachsemester.[11] Jeder Student erzeugt also eine Linie,[12] die dann stoppt, wenn der Student sich exmatrikuliert.
Zur Orientierung ist als Treppenfunktion das Soll von 30 Leistungspunkten (LP) pro Semester eingezeichnet, die bis zu den 180 LP ansteigt, die die Studierenden zum Abschluss ihres Bachelor benötigen. Da allerdings einige Pflichtveranstaltungen abgeschlossen werden müssen, kann es vorkommen, dass ein Student mehr als 180 Studienpunkte braucht, um seinen Abschluss zu machen. Diese „Spaghetti-Plots“ sind für BWL-er und VWL-er separat dargestellt. In beiden Studiengängen sieht man eine Gruppe von Studierenden, die in höheren Fachsemestern keine Studienleistungen mehr erbringen, was sich in sehr flachen Zuwachskurven äußert.

Eine zweite Möglichkeit besteht darin, den Kohortenverbleib über der Studiendauer abzutragen. Hierbei können verschiedene Zustände identifiziert werden: Der Studierende ist noch im Studiengang eingeschrieben, er hat ihn abgeschlossen, er hat den Studiengang gewechselt oder er hat sich ohne Abschluss exmatrikuliert. Für eine Untersuchung auf Basis von administrativen Daten einer Universität kann der letzte Zustand nicht weiter differenziert werden. Vor demselben Problem standen auch die Untersuchungen von Berens u. a. (2019) für Wuppertal, Danilowicz-Gösele u. a. (2017) für Göttingen und Fleischer u. a. (2019) für Duisburg/Essen.

[11] Basis sind die vier Kohorten 2010/11 bis 2013/14
[12] Zur besseren Sichtbarkeit wurde die Sprungfunktion durch eine glatte Interpolation ersetzt.

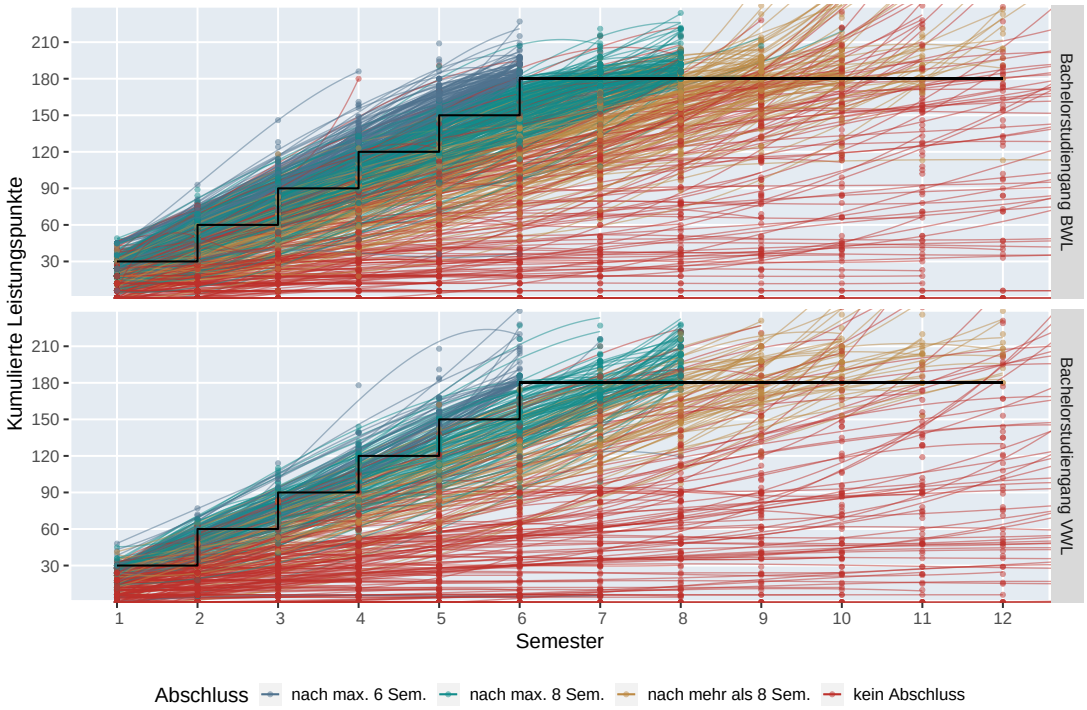


Abbildung 1.3: Entwicklung der individuell erreichten Studienpunkte (Kum_LP) nach Semester (Oben BWL, unten VWL). Der empfohlene Verlauf ist in schwarz als Referenz abgebildet.

Die Exmatrikulation ohne Abschluss bedeutet nicht in jedem Fall einen Austritt aus der tertiären Ausbildung, vgl. Heublein u. a. (2017). Es bestehen diverse Möglichkeiten im universitären System zu verbleiben:

- Es wird lediglich die Universität gewechselt, aber das Studium wird in demselben Studiengang fortgesetzt.
- Es wird die Universität und der Studiengang gewechselt.
- Das Studium wird unterbrochen (wegen Finanzierung, Kinderbetreuung, Krankheit etc.) und später fortgesetzt.

Bei einer retrospektiven Erhebung unter Studienabbrechern können diese Sachverhalte erfragt werden, vgl. Heublein u. a. (2017). Allerdings besteht auch hier das Problem, dass längere Studienpausen nur mit einem großen zeitlichen Abstand zu erkennen sind. Weiterhin sind bei einem Wechsel des Studienorts Nonresponse-Probleme virulent, vgl. Basic und Rendtel (2007) und Schimpl-Neimanns (2008). Erst nach Realisation der im Hochschulstatistik-Gesetzes §3 festgelegten Merkmale (vgl. `https://www.gesetze-im-internet.de/hstatg_1990/index.html`) in einem für die Wissenschaft zugänglichen Register können Analysen an dieser Stelle differenzierter gestaltet werden.

Letztendlich ist die Ermittlung der Anzahl der Studienabbrecher auch eine definitorische Frage, was ein Studienabbruch ist, vgl. etwa Schröder-Gronostay (1999). Je nach Definition wird man

dabei zu unterschiedlichen Abbruchraten kommen. Allerdings ergibt sich bei weniger restriktiven Definitionen von Studienabbruch das Problem der Schätzung der Ausfallquote. Hier benutzt das DZHW eine synthetische Erfolgsquote auf Basis der Hochschulstatistik, die alle Abschlüsse in einem Studienfach in einem Jahr zu einem synthetischen Studienanfängerjahrgang ins Verhältnis setzt. Allerdings mittelt eine solche fachbezogene Quote über alle Universitäten. Über die Performance eines Studiengangs an einer bestimmten Universität ist damit noch nichts ausgesagt.

Aus der ökonomischen Sicht der betroffenen Fachbereiche stellt jeder Wechsel aus dem lokalen Studiengang ohne Abschluss einen Verlust dar, der bei der Vergabe der universitären Mittel negativ zu Buche schlägt. Aus diesem Grund und wegen der Restriktion auf die Prüfungsdaten einer Universität, benutzen wir hier eine restriktive Definition von Studienabbruch, als Exmatrikulation ohne Abschluss im gewählten Studiengang.[13]

Abbildung 1.4 stellt die Kohorte WS 2010/11 nach ihrem Verbleib im Studiengang dar. Der sich über die Semester vergrößernde dunkel blaue Bereich (rechts oben) zeigt den Anteil derjenigen, die im Studiengang ihren Abschluss gemacht haben. Im unteren roten Teil der Graphik ist der Anteil derjenigen Studierenden dargestellt, die noch im jeweiligen Studiengang eingeschrieben sind. Im mittleren Bereich sind die Exmatrikulierten (Farbe türkis) sowie die Studienfachwechsler an der FU (Farbe orange) aufgeführt. Bei den Exmatrikulierten fehlt die Information, ob lediglich die Universität gewechselt wird oder ob es sich um einen Studienabbruch handelt. Wiederum wird die Entwicklung für die BWL (obere Grafik) und die VWL (untere Grafik) separat dokumentiert.

Zunächst fällt auf, dass schon im ersten Semester ein Schwund von 10 Prozent des Bestands der Kohorte eingetreten ist. Dahinter verbergen sich meist verspätete Zulassungen in anderen Studiengängen oder Universitäten, für die die Studierenden eine höhere Präferenz haben. Weiterhin zeigt sich, dass der Bestand an Studierenden mit einem Abschluss in 8 Semestern (= Regelstudienzeit + 2 Semester) bei den BWL-ern mit 42% deutlich größer ist als bei den VWL-ern mit 28%. Allerdings ist der Bestand an Langzeitstudenten mit mehr als 12 Semestern, die noch im Studiengang verbleiben, bei beiden Studiengängen mit 7% bzw. 5% in etwa gleich. Insgesamt ist der Anteil der VWL-er mit einem Wechsel des Studiengangs an der FU deutlich größer als bei der BWL. Dies gilt auch für die Exmatrikulationen.

An der HU erhält man für die gleiche Kohorte ähnliche Stabilitätsziffern: Dort erreichen 39% der BWL-Studenten bis zum 8. Semester einen Abschluss. Allerdings ist hier die Erfolgsrate bei der VWL mit 36% im Vergleich zu 28% an der FU doch deutlich höher. Der Anteil der Langzeitstudenten fällt mit 7% für beide Studiengänge wie an der FU aus.

Allerdings lassen sich in Abbildung 1.4 die universitätsinternen Studienfachwechsler (Bereich ocker oberhalb der Bestands im Studiengang) direkt identifizieren. Für die BWL ist dieser Anteil sehr gering und auch bei der VWL ist der kumulative Anteil der Fachwechsler bis zum

[13] In dieser engen Definition wird auch ein Wechsel des Studienfachs als Abbruch gewertet.

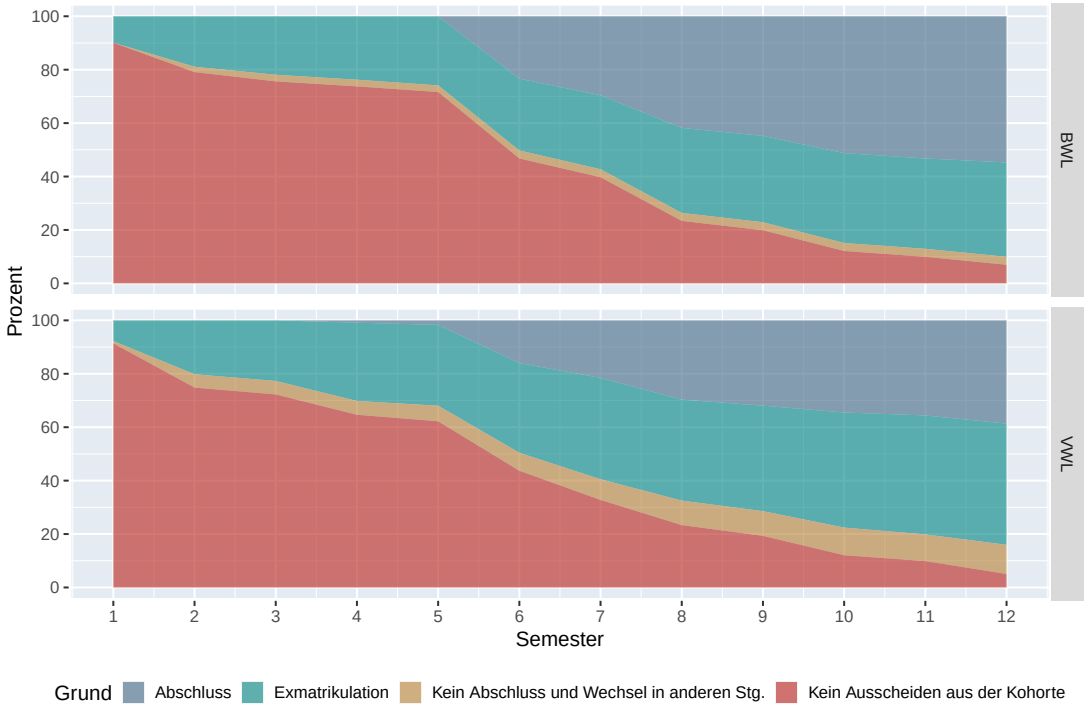


Abbildung 1.4: Entwicklung des Bestands der Kohorte WS10/11 nach Studiengang (Oben BWL unten VWL)

12 Fachsemester nicht größer als 10 Prozent des Ausgangsbestands bei Start des Studiums. Insgesamt scheinen Fachwechsel an derselben Universität keine große Rolle zu spielen. Studiengangwechsler, die aber die Universität wechseln, lassen sich in dem von uns verfolgtem Ansatz nicht identifizieren. Jedoch lässt die Auswertung der Hörsaalbefragung den Schluss zu, dass die meisten Studierenden nach einem vorzeitigen Ausscheiden aus dem Studiengang ein anderes Studium anstreben (84.3% der Studierenden mit Studienwechselneigung).

Eine weitere Darstellung der Studienleistung benutzt eine Gruppierung der Studierenden nach der Dauer bis zu ihrem Studienabschluss. In der folgenden Abbildung 1.5 wird für jedes Fachsemester die Summe der erworbenen Leistungspunkte über ein Boxplot dargestellt. In dieser Darstellung sind die Daten von vier Kohorten (WS 10/11 bis WS 13/14) zur Erhöhung der Fallzahlen kumuliert worden.

Die Darstellung differenziert nach BWL-ern (Oben) und VWL-ern (unten). Die Gruppe, die ihren Abschluss in der Regelstudienzeit von 6 Semestern schafft, erfüllt in jedem Semester im Mittel die in den Studienplänen geforderte Norm von 30 Leistungspunkten. Hier gibt es keine Unterschiede von BWL-ern und VWL-ern. Die nächste Gruppe ist definiert durch einen Studienabschluss bis zu 8 Semestern. Hier liegt die Studienleistung im 5. Semester etwas geringer, insbesondere bei den VWL-ern. Hier zeigen sich die Auswirkungen der Erasmus-Aufenthalte, die bevorzugt im 5. Semester absolviert werden. Wie man aus der Darstellung entnimmt, wird dies im 6. Semester teilweise wieder kompensiert. Dies gilt für beide Studiengänge. Wirklich dramatisch ist Studienleistung bei den Studierenden, die bis zum 8. Semester noch kei-

nen Abschluss gemacht haben. Hier ist die Studienleistung vom Studienstart an gleichmäßig sehr gering. Im Mittel werden nur 1 bis 2 statt der im Studienplan vorgesehenen 5 Module absolviert. Allerdings werden in dieser Darstellung nur die erreichten Leistungspunkte dokumentiert. Auch hier zeigen sich keine auffälligen Unterschiede zwischen den beiden Studiengängen.

Der Erwerb der Studienpunkte bei Studierenden ohne Abschluss bis zum 8. Semester liegt also mit ca. 6 bis 10 LP pro Semester deutlich unter der in der Studienordnung vorgesehenen Norm von 30 LP. Hier sind zwei Szenarien möglich. Diese Studierenden erwerben unterdurchschnittlich viele Leistungspunkte, weil sie viele Prüfungen nicht bestehen,[14] oder aber die Studierenden belegen gar nicht so viele Module. Dies kann mit unterschiedlicher Motivation geschehen. Zum einen kann ein Job während der Vorlesungszeit zu mangelnder Zeit für das Studium führen und damit eine Art Teilzeitstudium[15] erzwingen. Zum anderen gibt es die Strategie schwacher Studierender, sich erst mal auf wenige Veranstaltungen zu konzentrieren, die man dann nach und nach mit viel Vorbereitung auch abschließt.

Abbildung 1.6 zeigt die Studiensituation für die BWL an der HU. In der oberen Graphik sind die in den jeweiligen Fachsemestern erworbenen Leistungspunkte (LP) aufgeführt. Basis sind Studierende ohne Abschluss bis zum 8. Semester. Auch hier zeigt sich wie an der FU ein über alle Semester gleichmäßig geringer Studienerfolg von ca 10 LP, was etwa zwei abgeschlossenen Modulen entspricht. Allerdings nimmt auch hier der Studienerfolg nach dem 8. Semester noch einmal deutlich ab. Dies war auch an der FU zu beobachten. Weiterhin stehen insbesondere im ersten Studienjahr diesem geringen Leistungspunkteerwerb kaum nicht bestandene bzw. nicht angetretene Prüfungen gegenüber, vgl die beiden unteren Graphiken in Abbildung 1.6. Man kann es nicht anders sagen: Die Langzeitstudierenden gehen ihrem Studium nur mit geringer Intensität nach, und dies gilt von Beginn des Studiums an!

Bei der VWL ist diese Entwicklung ähnlich, vgl. Abbildung 1.7. Allerdings ist hier die Neigung Prüfungen nicht zu bestehen bzw. nicht anzutreten etwas größer als in der BWL. So sind am Ende des ersten Studienjahrs im Mittel 2 Prüfungen[16] nicht erfolgreich abgeschlossen worden, während es in der BWL im Mittel deutlich weniger als eine Prüfung war. Auch nach dem 8. Semester steigt in der BWL die Anzahl der nicht erfolgreich abgeschlossenen Prüfungen nicht an. Anders in der VWL: Die Langzeitstudenten der VWL scheinen nach dem 8. Semester noch einmal einen – allerdings erfolglosen – Versuch zu unternehmen, zu einem Studienabschluss zu kommen.

[14] In diesem Punkt unterscheiden sich die FU und die HU. An der HU müssen sich die Studierenden vor jeder Prüfung aktiv zur Prüfung anmelden. An der FU sind Studierende mit Anmeldung zum Modul auch automatisch zur Modulprüfung angemeldet. Aus diesem Grund wird an der HU das Nichterscheinen nach Anmeldung zur Prüfung wie ein Nichtbestehen gewertet.

[15] Obwohl offiziell die Möglichkeit eines Teilzeitstudiums besteht, wird diese Alternative praktisch nicht genutzt.

[16] Die Mittelwerte belegen jeweils eine nicht bestandene und eine nicht angetretene Prüfung bis zum 2. Semester.

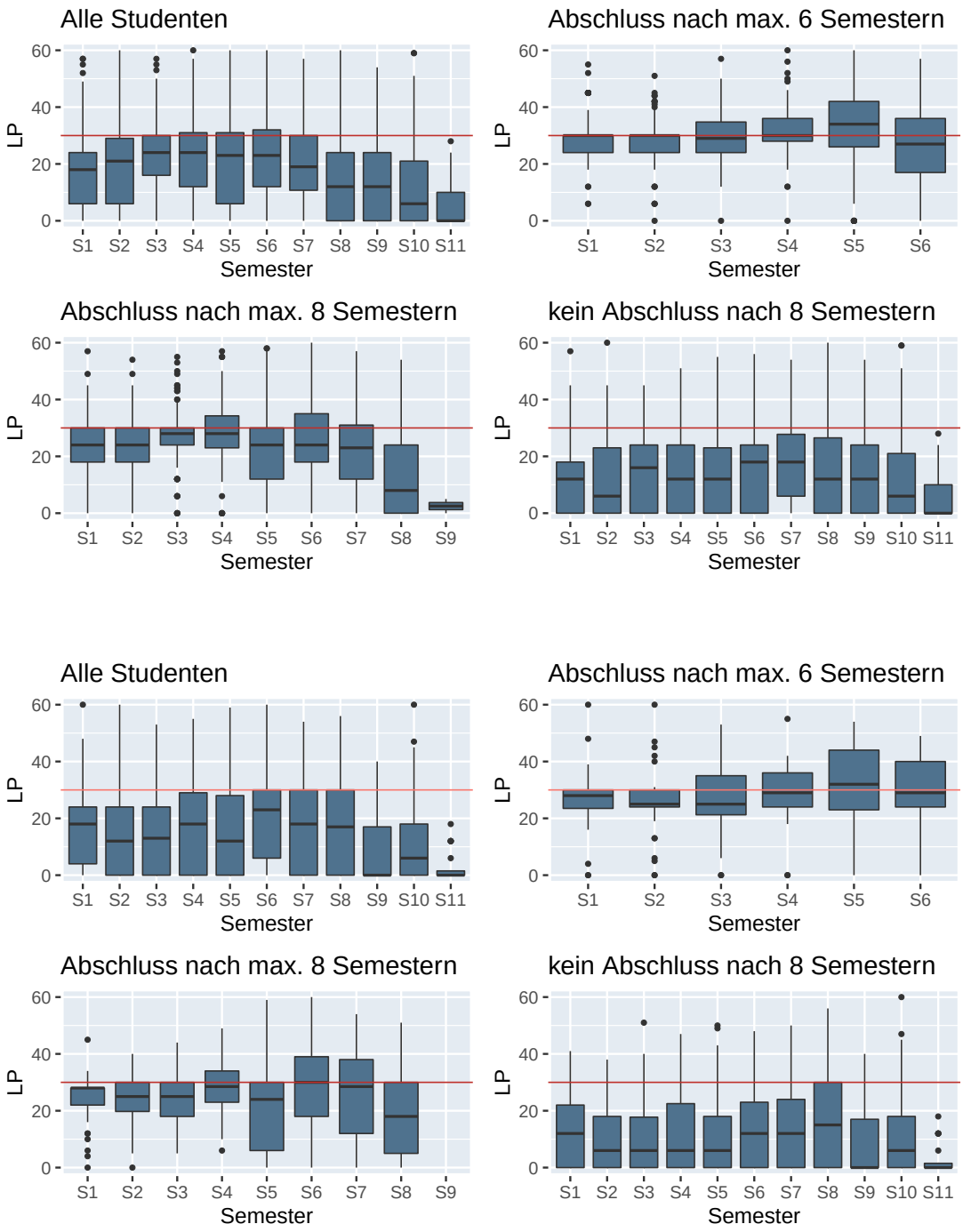


Abbildung 1.5: Erworbene Leistungspunkte nach unterschiedlichen Gruppen. (Oben BWL unten VWL). Kumulation über die Kohorten WS 10/11 bis WS 13/14

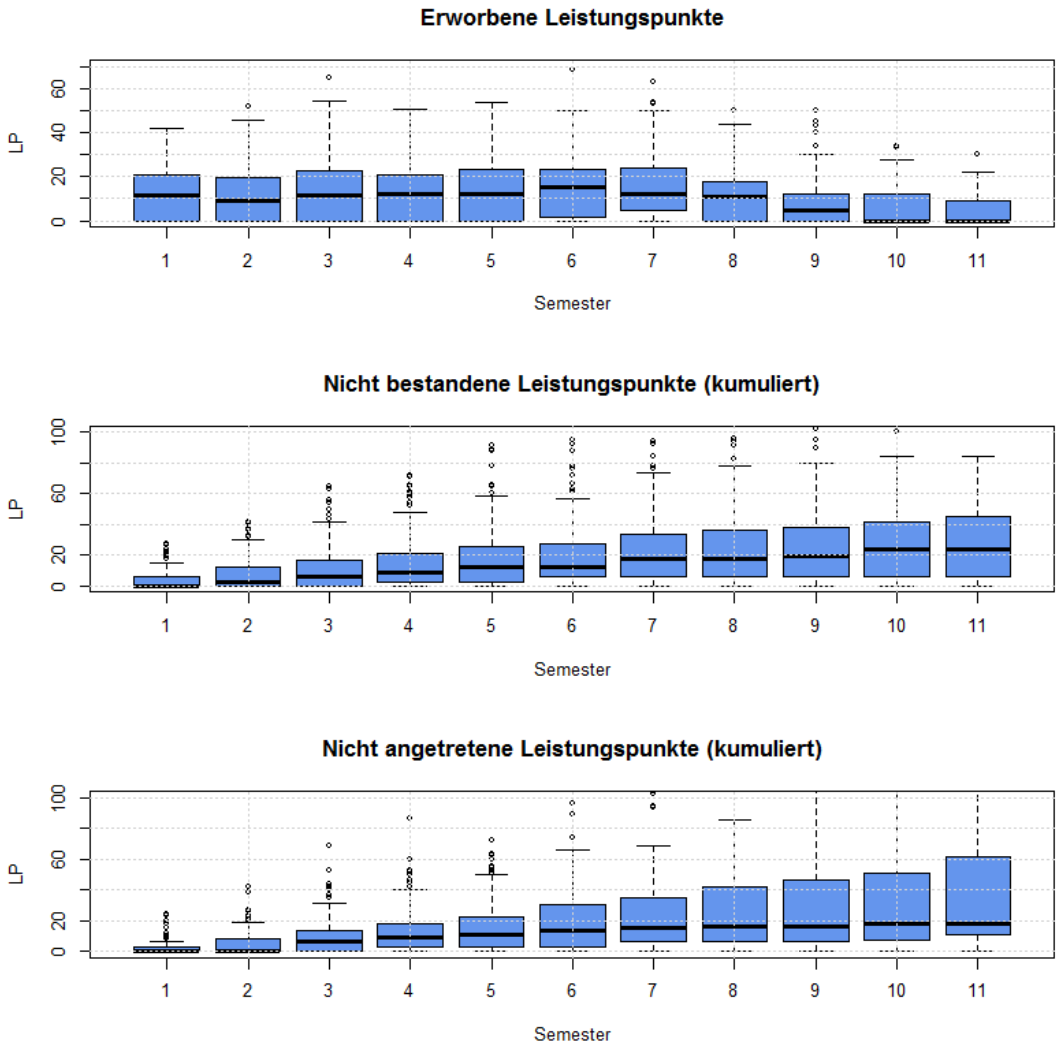


Abbildung 1.6: Erworbene Leistungspunkte, kumulierte Punkte von nicht bestandenen Prüfungen bzw. nicht angetretenen Prüfungen bei Studierenden ohne Abschluss bis zu 8. Semester (BWL HU). Kumulation über verschiedene Kohorten

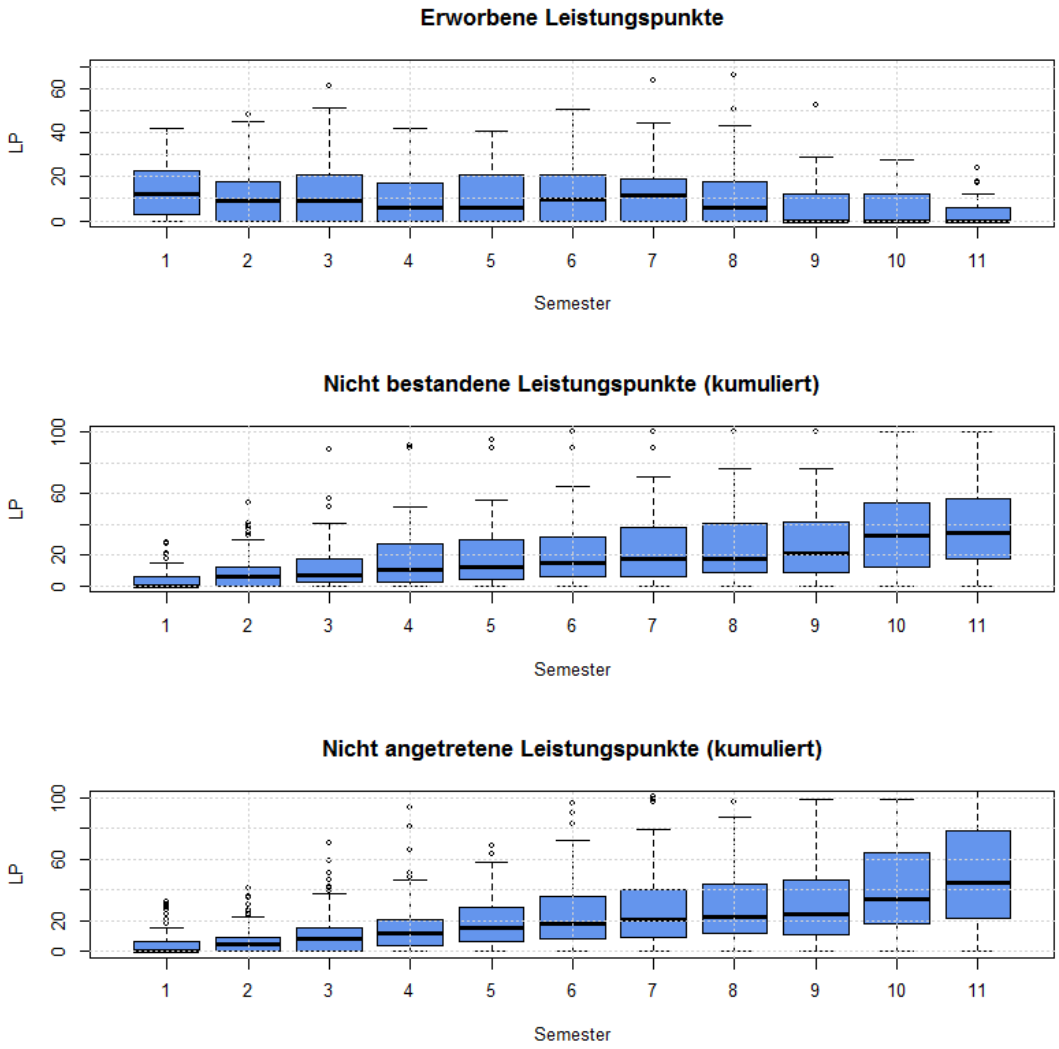


Abbildung 1.7: Erworbene Leistungspunkte, kumulierte Punkte von nicht bestandenen Prüfungen bzw. nicht angetretenen Prüfungen bei Studierenden ohne Abschluss bis zu 8. Semester (VWL HU). Kumulation über verschiedene Kohorten

1.4 Verzögerungen im Studienablauf

Die Studienpläne empfehlen für jeden Abschluss einen mehr oder weniger strengen Ablauf von Modul-Prüfungen vor. Beispielsweise sieht der BA VWL an der FU im ersten Semester das Modul „Mathematik“ vor, gefolgt von einer „Statistik für Wirtschaftswissenschaftler“ im zweiten Semester und einer „Schließenden Statistik“ im dritten Semester. Theoretisch ließe sich die Befolgung dieser empfohlenen Sequenzen anhand der Studienverläufe überprüfen. Doch dies scheitert in der Regel an der Vielfalt der sehr spezifischen Sequenzen. Etwas einfacher ist die Prüfung, wann welches Modul im wievielten Fachsemester abgeschlossen wurde, vgl. Hahm und Storck (2018). In der Regel treten Verspätungen von einem oder sogar zwei Jahren auf. Diese Verspätungen sind dem jährlichen Angebotsrhythmus der jeweiligen Module geschuldet. Wenn die Prüfung im Mathematik-Modul im ersten Semester nicht bestanden wurde, so muss man bis zum dritten Semester für einen nächsten Versuch warten.

Eine andere Art der Verzögerung ergibt sich aus dem geforderten Soll von 30 Leistungspunkten pro Semester. Defizite können hier durch nicht bestandene Modulprüfungen entstehen oder dadurch, dass Modulprüfungen überhaupt nicht angetreten werden. Im Extremfall wird das Studium nur noch sporadisch betrieben, so dass die Studierenden von Semester zu Semester immer weiter gegenüber dem 30 LP Leistungslimit zurückbleiben.

Die folgende Tabelle 1.2 gibt einen Überblick wie dynamisch der Prozess des Leistungsverzugs ist. Da im BA-Studium jedes Modul 6 Leistungspunkte (LP) erbringt, bedeuten die ausgewiesenen Kategorien einen Leistungsverzug von einer bis zu sechs und mehr Modulprüfungen.[17] Es zeigt sich, dass zwei Drittel der Studierenden ohne Leistungsverzug von einem Semester ins nächste Semester ohne weiteren Leistungsverzug kommen. Diese Studenten bleiben also im Studiensoll. Ein Viertel kann einen Verzug von einem Modul im kommenden Semester wieder kompensieren. Die Hälfte der Studierenden mit einem Verzug von 6 LP erleidet im kommenden Semester jedoch einen weiteren Verzug, meist um ein oder zwei Modulprüfungen (41% kumuliert). Dieses Ungleichgewicht impliziert einen Abwärtstrend im Verlauf des Studiums. Bei noch größeren Verzögerungen in der Studienleistung sieht es ziemlich hoffnungslos aus. 94% der Studierenden mit einem Verzug von mehr als 5 Modulen bleiben in dieser Kategorie. Bei Verzug von 4 Modulen tritt in zwei Drittel aller Fälle eine weitere Verschlechterung ein.

Die hier aufgezeigte Dynamik belegt deutlich, dass größere Verzögerungen im Studienablauf nicht mehr aufgeholt werden sondern sich im Gegenteil weiter vergrößern.

Auf Grund dieser Erkenntnis haben wir untersucht, in wie weit sich die Leistungspunkte, welche zu Beginn des Studiums erworben werden, durch a-priori bekannte Faktoren[18] erklären lassen. Insbesondere interessieren wir uns hier für den Einfluss der Hochschulzugangsberechtigungsnote (HZB-Note, Bezeichnung ”Abi”) sowie soziodemografischer Hintergrundvariablen.

[17] Es kann durch Wechsler des Fachs, der Prüfungsordnung oder der Universität in Einzelfällen auch zu Verzügen zwischen den 6-Punkt-Schritten kommen.

[18] Eine Liste der zur Verfügung stehenden Variablen ist in Appendix 1.8.1 gegeben

	Verzug im Folgesemester						
$t_0 \setminus t_1$	kein	1-6 LP	7-12 LP	13-18 LP	19-24 LP	25-30 LP	>30 LP
kein	**0.68**	0.08	0.14	0.05	0.02	0.02	0.01
1-6 LP	**0.27**	0.22	**0.27**	**0.14**	0.04	0.02	0.04
7-12 LP	0.14	0.08	0.20	0.27	0.14	0.06	0.11
13-18 LP	0.10	0.05	0.08	0.19	0.19	0.15	0.24
19-24 LP	0.05	0.01	0.05	0.07	0.13	0.20	0.49
25-30 LP	0.03	0.02	0.02	0.05	0.08	0.13	**0.67**
>30 LP	0.01	0.00	0.01	0.01	0.01	0.02	**0.94**

Tabelle 1.2: Übergangsmatrix des Verzuges in den ersten 6 Semestern

Die Variablenselektion erfolgte automatisch unter Optimierung des BIC. Das resultierende Modell ist in Tabelle 1.3 dargestellt. Das zugehörige Bestimmtheitsmaß beträgt nur 13% und das adjustierte Bestimmtheitsmaß lediglich 9%. Somit bleiben 87% des Leistungspunkterwerbs unerklärt.

Es steht demnach kein starker Prädiktor für den Studienerfolg zur Verfügung. Insbesondere zeigt dies, dass die HZB-Note nur wenig über die Studieneignung aussagt, wenngleich sie besser geeignet ist als nahezu alle anderen a-priori bekannten Variablen[19]. Man kann zusammenfassend sagen, die HZB-Note ist der beste Indikator von vielen schlechten Prädiktoren.

Tabelle 1.3: Regression des kumulierten Erwerbs von LP in den ersten zwei Semestern

	Schätzer	Pr(>|z|)
Konstante	54.80	**0.00**
Studium VWL	-2.57	0.38
Abi 1.6 - 2.0	-4.08	0.22
Abi 2.1 - 2.5	-6.27	0.11
Abi 2.6 - 3.0	-17.10	**0.00**
Abi 3.1 - 3.5	-8.15	0.28
Abi >3.5	-41.50	**0.03**
Abitur nicht in Berlin	-7.34	**0.01**
kein Nebenerwerb	-4.75	0.07

Wir versetzten jetzt den Betrachtungszeitpunkt auf den Beginn des zweiten Semesters und stellen uns wieder die Frage, wie gut wir den Erwerb von LP in den folgenden zwei Semestern vorhersagen können. Zu diesem Zeitpunkt stehen erste Verlaufsdaten zur Verfügung, insbesondere sind die im ersten Semester erworbenen Leistungspunkte bekannt, welche wir als Kovariate ergänzen.

Das resultierende Modell weist ein Bestimmtheitsmaß von 45% sowie ein adjustiertes Bestimmtheitsmaß von 43% auf, erklärt somit den Leistungspunkterwerb wesentlich besser als

[19] Die Referenzkategorie ist hier durch eine HZB-Note bis zu 1.5 gegeben. Erst ab einer HZB-Note, die schlechter als 2.6 ist, werden Unterschiede im Erwerb der Studienpunkte als signifikant erkannt

Tabelle 1.4: Regression des kumulierten Erwerbs von LP in den Semestern 2 und 3

	Schätzer	Pr(>|z|)
Konstante	29.30	**0.00**
LP Sem 1	1.10	**0.00**
Studium VWL	-10.80	**0.00**
Abi 1.6 - 2.0	-3.09	0.31
Abi 2.1 - 2.5	-8.13	**0.02**
Abi 2.6 - 3.0	-12.10	**0.01**
Abi 3.1 - 3.5	-2.41	0.72
Abi >3.5	-15.80	0.34
Abitur nicht in Berlin	-4.10	0.09
kein Nebenerwerb	-3.37	0.16

ein Modell ohne Verlaufsdaten. Für jeden erworbenen Leistungspunkt im ersten Semester steigt die mittlere Anzahl LP in Semester zwei und drei um etwa einen LP. Ein Student, der also die planmäßigen 30 LP im ersten Semester erworben hat und für den keine der im Modell enthaltenen Dummyvariablen zutrifft, erwirbt über die nächsten zwei Semester im Mittel weitere 62.3 LP. Eine HZB-Note schlechter als 1.5 und ein nicht in Berlin erworbenes Abitur verschlechtern den Leistungspunkterwerb dagegen tendenziell. Überrascht wurden wir von dem negativen Effekt[20], die Hochschulzulassungsberechtigung nicht in Berlin erworben zu haben. Als mögliche Erklärung vermuten wir, dass die Wahl des Wohnortes eventuell wichtiger war, als der Studiengang und die Universität.

1.5 Studienwechselneigung und Studienabbruch

In der Studieneingangsphase konkretisieren sich die Erwartungen der Studierenden an ihr Studium. Aber auch die Studierfähigkeit der Schulabgänger zeigt sich nach den ersten Prüfungen. In dieser Studienphase werden bereits erste Erwartungen hinsichtlich des Abschlusses des Studiums formuliert. In dieser Hinsicht ist der Beginn des zweiten Semesters, wenn alle Prüfungsergebnisse aus dem ersten Semester vorliegen, ein geeigneter Zeitpunkt für die Erhebung einer Studienwechselneigung.[21] Allerdings ist nicht klar, inwieweit eine hohe Studienwechselneigung auch ein guter Indikator für einen realen Studienabbruch ist.

In der Literatur wurde diese Fragestellung mit unterschiedlichen Ergebnissen diskutiert. Gold (1988) und Bean (1982) berichten mit stichprobenbasierten Untersuchungen von einem Zusammenhang zwischen Studienwechselneigung und Studienabbruch, während Georg (2008) mit einer retrospektiven Querschnittserhebung diesen Zusammenhang als Spekulation bewertet. Die Untersuchung von Brandstätter u. a. (2006) findet für österreichische Studenten auf Basis einer Stichprobenauswahl keinen empirischen Zusammenhang zum Studienabbruch. Diese

[20] Wegen der schwachen Signifikanz ist dieser Effekt jedoch wenig abgesichert.

[21] Diejenigen Studierenden, die bereits nach dem ersten Semester das Studium abgebrochen haben, können allerdings nicht berücksichtigt werden.

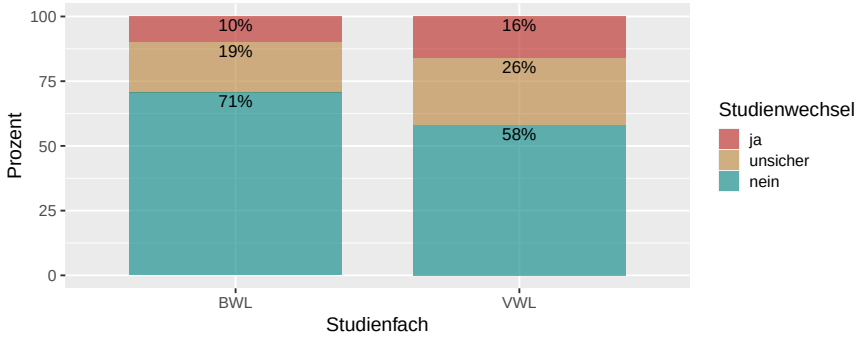


Abbildung 1.8: Studienwechselneigung nach Studiengang BWL und VWL

fachübergreifenden Studien wurden von Fleischer u. a. (2019) kritisiert, da Hinweise vorlagen, dass das Abbruchsverhalten zwischen den Studiengängen variiert. In einer fachspezifischen Untersuchung von naturwissenschaftlichen und ingenieurwissenschaftlichen Studiengängen untersuchen Sie daher das Abbruchsverhalten in der Studieneingangsphase, konkret Exmatrikulation bis zum Ende des zweiten Semesters. Die Rekrutierung an dieser freiwilligen Befragung erfolgte ebenfalls über Lehrveranstaltungen.[22] Hierbei wird die Wechselneigung am Ende des ersten Semesters erfragt, so dass der Prognosezeitraum bis zum Ende des zweiten Semesters lediglich 6 Monate beträgt. Für diesen Prognosezeitraum stellen Fleischer u. a. (2019) in allen Studiengängen einen signifikanten Einfluss der Wechselneigung auf den Studienabbruch fest.

In dieser Studie soll die Studienwechselneigung und ihre Prognosekraft über einen deutlich längeren Zeitraum überprüft werden. Im Rahmen unserer Studie war dies ein Zeitraum von 3.5 Jahren (= Abschluss 8. Fachsemester).[23] Zusätzlich wurden jedoch noch weitere Hintergrund- und Motivationsmerkmale im Rahmen einer multiplen Logitanalyse verwendet. Die Liste aller verwendeten Merkmale ist im Anhang A wiedergegeben. Aus prinzipiellen Gründen wurde das Merkmal Geschlecht in allen Analysen beibehalten.[24]

Die Studienwechselneigung und mögliche Ursachen wurden im Rahmen der Hörsaalbefragung explizit erfragt. Die Formulierung lehnte sich an die Querschnittsbefragung von Georg (2008) an: „Bedenken Sie einen Wechsel des Studiengangs?“ mit den Antwortkategorien „ja“, „unsicher“ und „nein“. Bereits bei Beginn des zweiten Semesters zeigte sich bei einem großen Teil der Studierenden doch Zweifel, ob sie den Studiengang zu Ende führen, vgl. Abbildung 1.8. Genau ein Drittel der 315 Antwortenden war sich nicht mehr sicher, ob sie ihr Studium zu Ende führen. Hierbei zeigten sich deutliche Unterschiede zwischen den Studienrichtungen BWL und VWL.

[22] Die hohe Ausschöpfungsquote von 71% wurde mit einer Bezahlung von 150 Euro plus 3 Leistungspunkten förmlich erkauft.

[23] Für die hier untersuchte Kohorte, die im SS 2016 im 2. Fachsemester war, war im WS19/20 dem Zeitpunkt der Abfassung des Manuskripts, im 9. Fachsemester. Die Angaben für höhere Fachsemester lagen also noch nicht vor.

[24] Allerdings erweist sich die Unterscheidung nach Männern und Frauen sowohl hinsichtlich der Studienwechselneigung als auch hinsichtlich des Studienabbruchs als nicht signifikant.

Bei der Analyse der Studienwechselneigung wird in der Literatur zwischen zwei Merkmalsgruppen unterschieden. Die Bedingungsfaktoren beziehen sich auf Einflüsse vor Beginn des Studiums wie Herkunft, Elternhaus, Belastung durch Nebenerwerbstätigkeit oder mangelnde Information über den Studiengang. Eine abgeschwächte Studienwechselneigung wird für die folgenden Merkmale erwartet:

- Gute kognitive Eingangsvoraussetzung (Hinneberg, 2003), die über die Hochschulzulassungsnote operationalisiert wird.
- Weiblich (Heublein u. a., 2017)
- Akademischer familiärer Hintergrund (Sarcletti und Müller, 2011)
- Abgeschlossene Berufsausbildung (Heublein u. a., 2017)

Dagegen bezieht sich die Merkmalsgruppe „Motivation“ direkt auf den Studiengang. Hierzu gehören Leistungsprobleme, mangelnde Motivation und negativ wahrgenommene Studienbedingungen. Unser Fragebogen ging auf mögliche Leistungsprobleme und Probleme bei der Realisation des Studiums detailliert ein, vgl die Dokumentation im Anhang A:

- Leistungspunkte (LP bzw. ECTS):
  - Erreichte LP im ersten Semester
  - Geplante LP
  - Differenz geplante - erreichte LP
- Prüfungen: ungeplante, nicht bestandene Prüfungen sowie geplante, nicht bestandene Prüfungen[25]
- Sicherheit bei der Wahl des Studiengang (5 stellige Likert Skala mit den Stufen „sehr sicher“ (5) bis „sehr unsicher“ (1))
- Ob der Studiengang die erste Wahl war
- Veränderung der Studienmotivation (5 stellige Likert Skala mit den Stufen „Starke Zunahme“ (5), „keine Veränderung“ (3) und „starke Abnahme“ (1).
- Wahl des Studiums aufgrund eines speziellen Schwerpunkts.
- Ausreichendes Modulangebot im Interessengebiet.

Da hier insbesondere die Unterschiede zwischen den Studiengängen BWL und VWL beleuchtet werden sollen, sei hier auf die unterschiedliche Bedeutung der Motivationsvariable „Wahl des Studiengangs wegen eines Studienschwerpunkts (ja/nein)“ verwiesen. Abbildung 1.9 zeigt,

[25] Da Wiederholungsprüfungen am FB Wirtschaftswissenschaft nur nach vorherigen Nicht-Bestehen abgelegt werden können, fallen einige Studierenden mit Absicht durch die Prüfung um am späteren Wiederholungstermin teilnehmen zu können. Dies geschieht in der Absicht, den Prüfungszeitraum um mehrere Monate zu vergrößern.

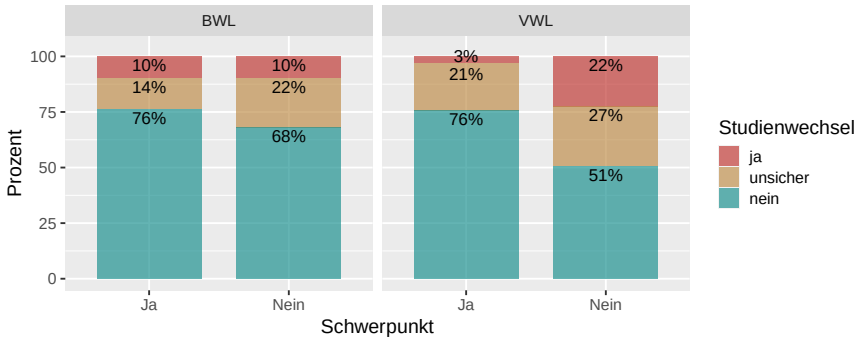


Abbildung 1.9: Studienwechselneigung nach Studiengang BWL und VWL. Interaktion nach Wahl des Studiengangs wegen eines Studienschwerpunkts (ja/nein)

dass für den Studiengang BWL die Bedeutung eines speziellen Studienschwerpunkts nur eine untergeordnete Rolle für das Studienabbruchempfinden hat. Beim Studiengang VWL liegen die Dinge jedoch anders. Wer hier keine feste Bindung an einen Studienschwerpunkt hat, schließt schon nach dem ersten Semester in 50% aller Fälle eine Studienwechsel nicht aus. Dies ließe sich im Sinne eines Verlegenheitsstudiums[26] interpretieren. In der BWL scheint dagegen eine klare Perspektive zu herrschen. Das häufige Vorurteil, die hohe Abbruchquote in der VWL läge an dem häufigen Wechsel in die BWL, können wir in unseren Daten nicht beobachten. Die Wechsel zwischen den Studiengängen BWL und VWL an der FU liegt in der betrachteten Kohorte im einstelligen Bereich und finden bidirektional statt. Insgesamt wechselte lediglich ein VWLer mehr in die BWL als andersherum.

Bei der Beurteilung der Unterschiede zwischen den Studiengängen BWL und VWL sollte auch die unterschiedliche Zusammensetzung hinsichtlich Note und Geschlecht berücksichtigt werden. Aufgrund der hohen Studiennachfrage ist der Leistungsdurchschnitt bei den BWL-ern deutlich günstiger (Median Abiturnote 1.8 (BWL) vs. 2.1 (VWL)). Weiterhin ist der Frauenanteil bei den BWL-ern mit 61% deutlich höher als bei den VWL-ern mit 35%. Aus diesem Grund müssen beide Merkmale bei einer multiplen Analyse mit berücksichtigt werden.

Nachfolgend sollen nun der simultane Einfluss die verschiedenen Merkmale auf die Variable Y = Studienwechselneigung mit einem ordinalen Logitmodell [27] (Fahrmeir u. a., 2013) geschätzt werden. Die Modellselektion wurde mittels schrittweiser Variablenselektion durchgeführt. Es wurde mit dem vollen Modell gestartet (siehe Appendix 1.8.1) und in jeder Iteration konnten Kovariaten sowohl entfernt, als auch hinzugefügt werden. Da diese automatische Modellselektion grundsätzlich in der Kritik steht, zu einer Überanpassung zu führen, wurde anstelle des häufig verwendeten Akaike-Informationskriterium (AIC) das strengere Bayessche Informationskriterium (BIC) optimiert. In Tabelle 1.5 sind die Koeffizientenschätzer des nach BIC optimalen Modells dargestellt. Es ergeben sich drei signifikante Variablen: Die erworbenen

[26] Eine etwas positivere Interpretation wäre eine geringe Vertrautheit mit den Studieninhalten.

[27] Die Parametrisierung ist durch $\log(\frac{P(Y_i \leq j)}{P(Y_i > j)}) = \kappa_j + x_i'\beta$ gegeben. D.h. positive Steigungskoeffizeiten zeigen eine Vergrößerung der Studienwechselneigung an, während negative Koeffizienten eine Verringerung des Studienwechselrisikos anzeigen.

Leistungspunkte des ersten Semesters, der Rang des Studiengangs (erste, zweite, dritte Wahl) sowie die Sicherheit, mit der das Studium gewählt wurde. Interessant ist, die anderen Merkmale zu Erfahrungen mit dem Studiengang, z.B. ein Indikator, ob sich gewisse Vorstellungen über den Studiengang negativ bestätigt, haben keinen signifikanten Einfluss auf die Studienwechselneigung. Auch der Einfluss der VWL auf die Studienwechselneigung liegt unterhalb der üblichen Signifikanzniveaus. Der große Wert des Koeffizienten für die Nennung des Studiengangs als dritte Wahl zeigt an, dass fast alle Studierenden mit dieser Einschätzung eine Studienwechselneigung äußern.

Tabelle 1.5: Geschätze Koeffizienten des ordinal-logistischen Modells zur Erklärung der Wechselneigung

Variable	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Fehler	t-Wert	$\Pr(>\|t\|)$
StgangVWL	0.684	1.982	0.368	-1.860	0.063
geschlm	0.243	1.275	0.350	-0.693	0.488
Sem_1_LP	-0.049	0.952	0.018	2.666	**0.008**
stud_rank3	3.229	25.253	0.859	-3.757	**0.000**
stud_rank2	1.069	2.911	0.377	-2.837	**0.005**
stud_sich	-1.095	0.335	0.215	5.086	**0.000**
studvorst_verneg. bestätigt	-0.272	0.762	0.567	0.480	0.631
Schwellenwerte	$\hat{\kappa}$		Std. Fehler		
ja|unsicher	2.546		0.973		
unsicher|nein	4.997		1.021		

Die Betrachtung der Prädiktionsgüte des ordinal-logistischen Modells zeigt, dass es nur schlecht in der Lage ist, zwischen den Ausprägungen „unsicher“ und „ja“ zu diskriminieren. Aus diesem Grund fassen wir diese beiden Kategorien in der neuen Variable „wechselneigung Bin“ wie folgt zusammen:

$$\text{wechselneigung Bin}_i = \begin{cases} 0 & \text{für Wechselneigung Ordinal}_i = \text{nein} \\ 1 & \text{sonst.} \end{cases}$$

Für diese neu definierte Wechselneigung führen wir erneut eine Variablenselektion nach BIC aus. Hierbei verwenden wir das gleiche Verfahren und die gleichen Variablenbasis wie schon für das ordinale Modell. Das resultierende Modell ist in Tabelle 1.6 dokumentiert. Die *area under the curve*(AUC) beträgt 0.82, was auf einen guten Modellfit hindeutet.

Die Interpretation der Koeffizienten ist hier dieselbe wie bei dem ordinalen Logitmodell von Tabelle 1.5. Auffällig ist hier wieder der hohe Wert für die Bezeichnung des Studiengangs als dritte Wahl. Hier müssen fast alle Studierenden mit dieser Nennung einen Wechsel des Studiums genannt haben[28]. Dies erscheint auch plausibel. In diesem Modell ist die Klassifika-

[28] Wegen der geringen Anzahl der Nennungen ist der Schätzwert für den Effekt allerdings nicht signifikant.

Tabelle 1.6: Geschätze Koeffizienten des binär-logistischen Modells zur Erklärung der Wechselneigung

Variable	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Fehler	z-Wert	Pr(>|z|)
Intercept	4.312	74.624	0.998	4.321	**0.000**
StgangVWL	0.489	1.631	0.378	1.295	0.195
geschlm	0.137	1.147	0.363	0.379	0.705
Sem_1_LP	-0.058	0.943	0.019	-3.092	**0.002**
stud_rank3	18.411	–	1124.255	0.016	0.987
stud_rank2	0.845	2.328	0.394	2.145	**0.032**
stud_sich	-1.151	0.316	0.235	-4.904	**0.000**

tionsmatrix in Tabelle 1.7 deutlich günstiger. Insgesamt werden in diesem Modell 77.8 Prozent der Studierenden hinsichtlich ihrer Studienwechselneigung richtig eingeschätzt.

Tabelle 1.7: Klassifikationsmatrix der Wechselneigung aus Modell 1.6.

Prädiktion	Beobachtung	
	keine Wechselneigung	Wechselneigung
keine Wechselneigung	**120**	22
Wechselneigung	27	**52**

Im folgenden soll nun überprüft werden, ob die am Anfang des zweiten Semesters geäußerte Studienabbruchneigung ein zuverlässiger Indikator für den tatsächlichen Studienabbruch ist. Allein die Größenverhältnisse der beiden Teilmengen lassen erwarten, dass es noch zu deutlichen Verschiebungen im Verlaufe der nächsten 7 Semester kommt. Während 2/3 der Studierenden an Anfang des zweiten Semesters keine Studienwechselneigung zeigen, schaffen jedoch nur ca. 40 Prozent der Studierenden einen Abschluss bis zum 8. Semester.

Wir interessieren uns also nicht mehr für die Wechselneigung der betrachteten Studierenden als abhängige Variable sondern wir möchten ermitteln, in wie weit mit den Informationen zu Beginn des zweiten Semesters ein späterer tatsächlicher Wechsel bzw. ein Studienabbruch prognostiziert werden kann. Zu diesem Zweck generieren wir aus den administrativen Verlaufsdaten für die Teilnehmer der Hörsaalumfrage eine Binärvariable „Abbruch erfolgt“ nach folgender Logik: Wenn der Studierende zum Ende des achten Hochschulssemesters noch in seinem Studiengang eingeschrieben ist oder diesen erfolgreich abgeschlossen hat bzw. schon ein Masterstudium[29] aufgenommen hat, nimmt die Variable den Wert 0 an und 1 in den übrigen Fällen. Ob ein Abbruch/Wechsel erfolgte, modellieren wir durch ein binäres logistisches Regressionsmodell. Die Variablenselektion erfolgt wie bei den vorangegangen Modellen durch eine schrittweise Selektion unter Optimierung des BIC, die Liste der verwendeten Variablen ist

[29]Obwohl einzelne Prüfungsleistungen (z.B. die Korrektur der Abschlussarbeit steht noch aus) noch ausstehen wird bereits ein Masterstudium begonnen. In diesem Fall hat der Studierende bereits seinen Bachelorstudiengang verlassen diesen aber noch nicht erfolgreich abgeschlossen. Solch ein Verhalten entspricht aber eher einem Abschluss als einem Wechsel.

in Appendix 1.8.1 gegeben.

Das resultierende Modell ist in Tabelle 1.8 gegeben. Die in-sample Vertrauenswahrscheinlichkeit beträgt 0.74%. Zur Bestimmung dieses Gütemaßes wurde die Fläche unter der ROC-Kurve optimiert (AUC = 0.80), um einen geeigneten Grenzwert zu ermitteln. Beobachtungen mit einer Wahrscheinlichkeit unterhalb oder gleich dieses Grenzwertes werden mit 0 klassifiziert andernfalls mit 1.

Tabelle 1.8: Logistische Regression des Abbruchs/Studienwechsels bis einschließlich des achten Semesters

	$\hat{\beta}$	$\exp(\hat{\beta})$	Std. Fehler	z-Wert	Pr(>|z|)
Intercept	-0.907	0.404	0.512	-1.771	0.077
StgangVWL	0.505	1.657	0.395	1.280	0.200
geschlm	0.166	1.181	0.388	0.430	0.668
Sem_1_LP	-0.071	0.932	0.021	-3.371	**0.001**
stud_wechselja	2.027	7.592	0.537	3.778	**0.000**
stud_wechselunsicher	1.438	4.210	0.410	3.510	**0.000**

Eine geringe Zahl von Leistungspunkten im ersten Semester und die Äußerung einer Wechselneigung zu Beginn des zweiten Semesters vergrößern nachhaltig das Risikos eines Abbruchs. Quantitativ vergrößert die Antwort, dass man an einen Wechsel denkt, das Abbruchrisiko um den Faktor 7.6. Wer dagegen nur unsicher ist, ob er das Studium weiterführt, vergrößert gegenüber den Sicheren das Risiko nur um den Faktor 4.2. Das Verhalten ist also durchaus sensibel hinsichtlich der Deutlichkeit, mit der eine Wechselneigung ausgedrückt wird. Wer nur 12 statt der geforderten 30 LP erworben hat, vergrößert sein Abbruchrisiko um den Faktor $\exp((30 - 12) \times 0.07) = 5.4$. Treffen beide Ereignisse zu, so vergrößert sich bei der Nennung „Wechsel=ja“ und nur zwei abgeschlossenen Modulen das Abbruchrisiko um den Faktor $7.6 \times 5.4 = 41.06$. Damit kann schon am Beginn des 2. Semesters die Risikogruppe der späteren Studienabbrecher relativ genau abgegrenzt werden.

1.6 Schulische Leistungsindikatoren und Studienerfolg

Schulische Leistungsindikatoren nehmen eine herausragende Stellung bei der Vergabe der Studienplätze ein. Das weitaus größte Gewicht hat dabei die Note der Hochschulzugangsberechtigung (HZB-Note). 60 Prozent der Studienplätze werden in Berlin über die HZB-Note[30] vergeben. 20 Prozent werden über eine Warteliste vergeben, deren Reihung ihrerseits die Abiturnoten vor Eintritt in die Wartezeit widerspiegelt. Die verbleibenden 20 Prozent werden über Härtefallregelungen vergeben. Die Begründung für die starke Berücksichtigung der Abiturnote besteht in der „berechtigten Annahme einer höheren Wahrscheinlichkeit für einen Studienabschluss“, so Heublein u. a. (2017).

[30]Gewisse Abweichungen entstehen durch eine weitere Gewichtung mit der Mathematik-, Deutsch- oder Englischnote.

Die empirische Basis für diese Einschätzung beruht in der Mehrzahl der Fälle auf Korrelationsanalysen der Abiturnote mit der Abschlussnote an der Universität. In einer Metaanalyse untersuchen Trapmann u. a. (2007) 26 internationale Studien zum Zusammenhang von schulischer Leistung und Studiennote. Nach diesen Studien ist die Korrelation zwischen der Abiturnote und Studiennote in Deutschland am höchsten. Allerdings kann eine Studiennote nur bei Abschluss eines Studiums ermittelt werden. Die Anzahl der analysierten Studien, die sich lediglich auf den Abschluss des Studiums beziehen, ist mit 4 wesentlich kleiner.

Stellvertretend sei hier die Studie von Heublein u. a. (2017) genannt. Sie basiert auf einem Vergleich der Abiturnoten der Studienabbrecher und der Absolventen des Studiums. Danach liegt die Durchschnittsnote für Absolventen bei 2.3 während sie bei den Studienabbrechern bei 2.7 liegt, vgl. Abbildung 5.22 in Heublein u. a. (2017). Allerdings deuten die Ergebnisse auch auf eine erhebliche Streuung der schulischen Leistungsindikatoren und des Studienerfolgs. So sind unter den Studienabbrechern allein 16 Prozent mit einem Abiturdurchschnitt von besser als 2.0. Umgekehrt sind unter den Absolventen 12 Prozent mit einer Abiturnote schlechter als 3.0, vgl. Abbildung 5.23 in Heublein u. a. (2017).

Wir untersuchen im Folgenden den Zusammenhang zwischen der HZB-Note und dem Erfolg im Studium. Wir betrachten Erfolg als ein ein zweiteiliges Konzept: zum Einen das Erreichen des avisierten Abschlusses in angemessener Zeit, zum anderen die im Studium erzielten Noten. Dem ersten Teil haben wir uns bereits in Abschnitt 1.4 gewidmet. Konditioniert man auf die erreichten Leistungspunkte im ersten Semester und die individuelle Studienmotivation so ergibt sich kein zusätzlicher Effekt der Abiturnote auf den Abschluss des Studiums, vgl. Tabelle 1.8. Allerdings wirkt sich die Abiturnote auf die Anzahl der erreichten Leistungspunkte im ersten Semester aus. Inspiziert man jedoch die Koeffizienten für den Einfluss der Abiturnote auf die Leistungspunkte in Tabelle 1.3 so zeigt sich, dass der Einfluss von 1.0 bis zur Note 2.5 praktisch identisch ist. Dies ist aber der Bereich in dem der Numerus Clausus stark unterschiedliche Zugangschancen für das Studium setzt.

Dieser Zusammenhang von Abiturnote und dem Erfolg im ersten Studienjahr soll hier noch einmal bivariat über einen Scatterplot-Smoother[31] und separat für die BWL und die VWl mit Ergebnissen aus der HU (Kohorte WS 2014/15) dargestellt werden.[32] Die Abbildung 1.10 zeigt für die BWL-er praktisch keinen Einfluss der Abiturnote in einem breiten Bereich von den Bestnoten bis etwa zur Marke von 2.5. Erst für schlechtere Schulnoten zeigt sich im Mittel eine Tendenz zu einem geringeren Studienerfolg. Die Ergebnisse stimmen sehr gut mit der multivariaten Schätzung für die FU überein. Allerdings offenbart die separate Schätzung des bivariaten Effekts für die VWL und die BWL, dass in der VWL der Einfluss der Abiturnote auch bei den guten Schulnoten noch deutlich messbar ist. Diese Ungleichheit des Einflusses der Schulnoten zwischen den beiden Studiengängen wird auch bei den einzelnen Modulnoten auftreten. Analysiert man den Zusammenhang von Schulnoten und Studienerfolg im Sinne

[31] Dies ist eine nicht-parametrische Darstellung des bedingten Erwartungswerts für die abhängige Größe bei gegebener Abiturnote (LOESS Kurve)

[32] Ganz ähnliche Ergebnisse erhält man für die Kohorten WS2013/13 und WS2013/14.

klassischer Korrelationsmaße, so erhält man für die BWL $R^2 = 0.10$ und für die VWL $R^2 = 0.18$. Die unerklärte Residualvarianz ist also der bestimmende Faktor für den Erfolg in der Studieneingangsphase!

Bevor wir den Einfluss der HZB-Note[33] auf die bei Studienabschluss erzielten Noten untersucht werden, werfen wir einen Blick auf den Einfluss auf der Abiturnote auf einzelne Modulnoten.

Selbst im Fall einer erfolgreichen Teilnahme an einer Modulprüfung gibt es Gründe für Zweifel an einer hohen Korrelation von Modulnote und Abiturnote. Beispielsweise hat das im Modul abgefragte Wissen in der Regel nur wenig mit dem Schulwissen gemeinsam. Auch sind die Ergebnisse schriftlicher Klausuren, die am Fachbereich Wirtschaftswissenschaft die Regel sind, von gewissen Zufälligkeiten abhängig, wie zum Beispiel Koinzidenzen von Prüfungsstoff und individueller Vorbereitung.

Die Abbildungen 1.11 und 1.12 belegen diese geringe Korrelation mit diversen Streudiagrammen von Modulnoten und Abiturnote (HZB-Note) an der Wirtschaftswissenschaftlichen Fakultät der HU. Aus Gründen der Erhöhung der Fallzahl, aber auch aus Gründen der Anonymisierung wurden diese Ergebnisse über mehrere Kohorten (WS2010/11 bis WS 2015/16) gepoolt.

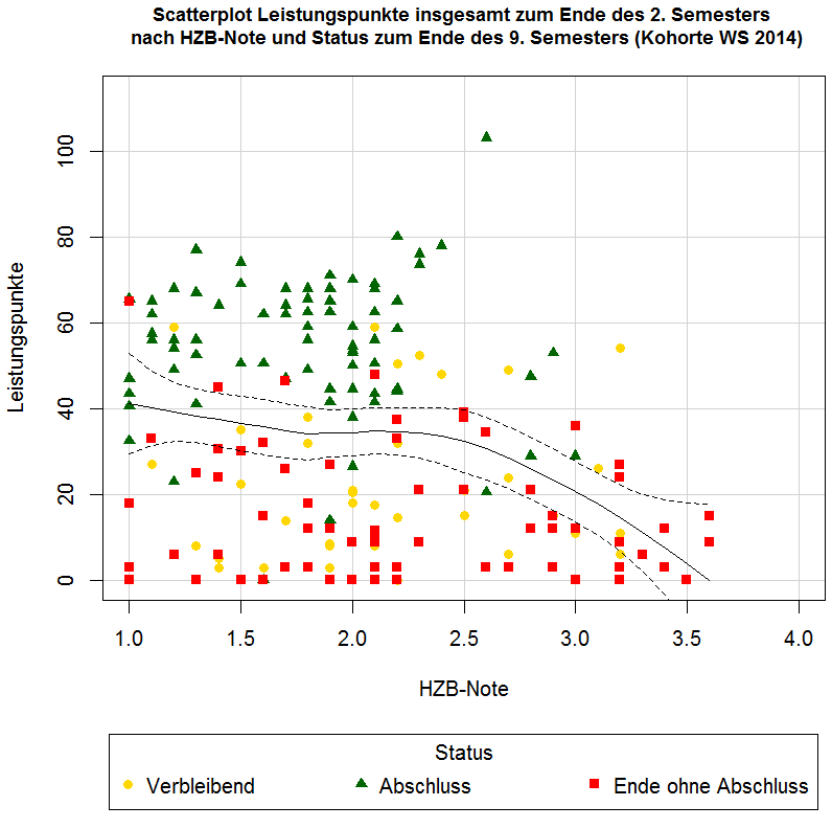
Die Streudiagramme werden durch einen Scatterplot Smoother ergänzt. Hier zeigt sich bei den BWL-ern ein systematischer Zusammenhang zwischen dem Abschluss des Mathematik I Moduls und der Schulnote. Wegen der Nähe der Mathematikausbildung zum Schulstoff erscheint dies wenig verwunderlich. Bei allen anderen Pflichtmodulen existiert jedoch gar kein Zusammenhang. Hervorstechend ist jedoch in erster Linie die hohe Residualvarianz, die alle systematischen Zusammenhänge überlagert.

Wechselt man von den BWL-ern zu den VWL-ern, so zeigen sich bei allen Modulen stärkere Zusammenhänge zwischen der Abiturnote und der Modulnote. Dies gilt selbst für Module, die sich wenig mit dem üblichen Schulstoff überlappen, wie zum Beispiel das Modul Recht. Viele Module im BA-Studium werden von BWL-ern und VWL-ern gemeinsam besucht.Auch die Prüfungsform ist bei beiden Studiengängen durchgängig eine schriftliche Klausur. Diese Differenzen im Studienerfolg können daher nur mit unterschiedlichen Lernkulturen[34] in diesen beiden Studiengängen erklärt werden. Trotzdem ist auch bei den VWL-ern die Residualvarianz die bestimmende Größe.

Es ist daher zu erwarten, dass die Bachelornote als Resultat dieser stark streuenden Modulnoten

[33] Die Analyse der Modulnoten erfolgt auf Basis der administrativen Daten der HU, die neben den Einzelnoten auch die Hochschulzugangsberechtigungsnote (HZB_Note) aus dem Zulassungsverfahren verfügbar macht. Für die FU standen in dem Projekt nur die erworbenen Leistungspunkte nicht jedoch die dabei erzielten Noten zur Verfügung. Auch werden an der FU die Daten von Zulassungsverfahren und Prüfungsadministration getrennt gehalten.

[34] Man könnte diesen Befund zunächst für einen verkappten Geschlechtseffekt halten, da an der HU der Männeranteil in der BWL 50% in der VWL aber 66% beträgt. Allerdings zeigt eine separate Auswertung nach Geschlecht einen unveränderten Befund.



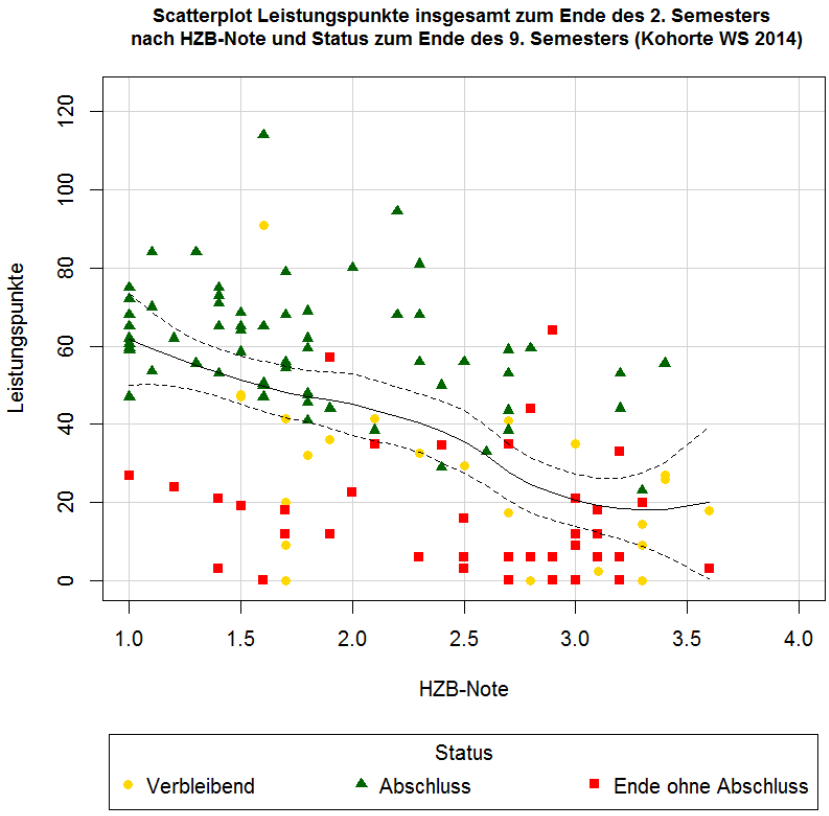


Abbildung 1.10: Der Einfluss der Abiturnote auf erreichten Leistungspunkte am Ende des ersten Studienjahrs (Kohorte WS 2014/2015, HU). Obere Grafik: BWL Untere Grafik: VWL Farben: Unterscheidung nach dem Studienstatus am Ende des 9. Semesters (Grün mit Abschluss, Rot ohne Abschluß exmatrikuliert, Gelb im Studiengang verbleibend)

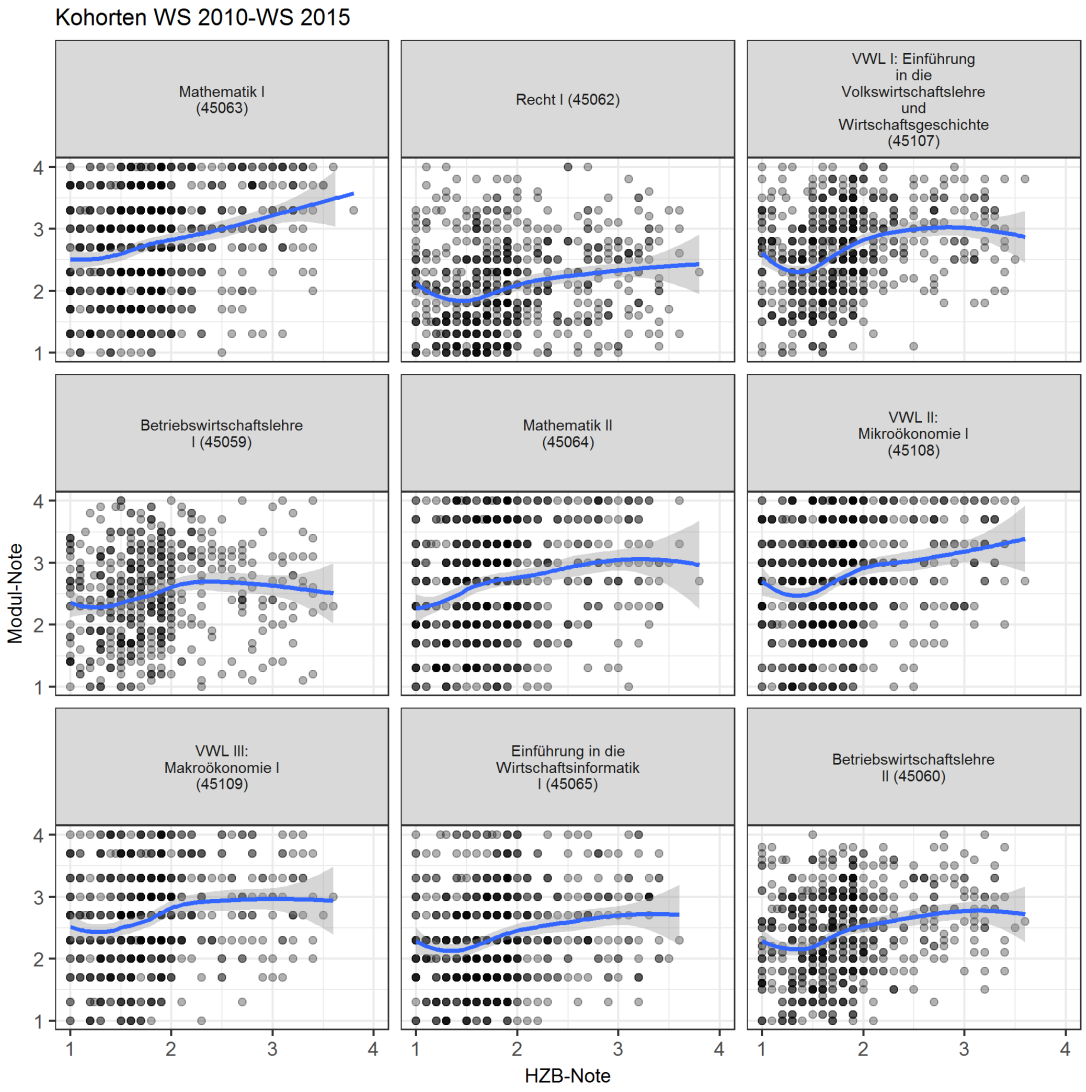


Abbildung 1.11: Erreichte Modulnoten nach Abiturnote (HZB-Note) in BWL. Scatterplot Smoother (LOESS) mit Konfidenzintervall

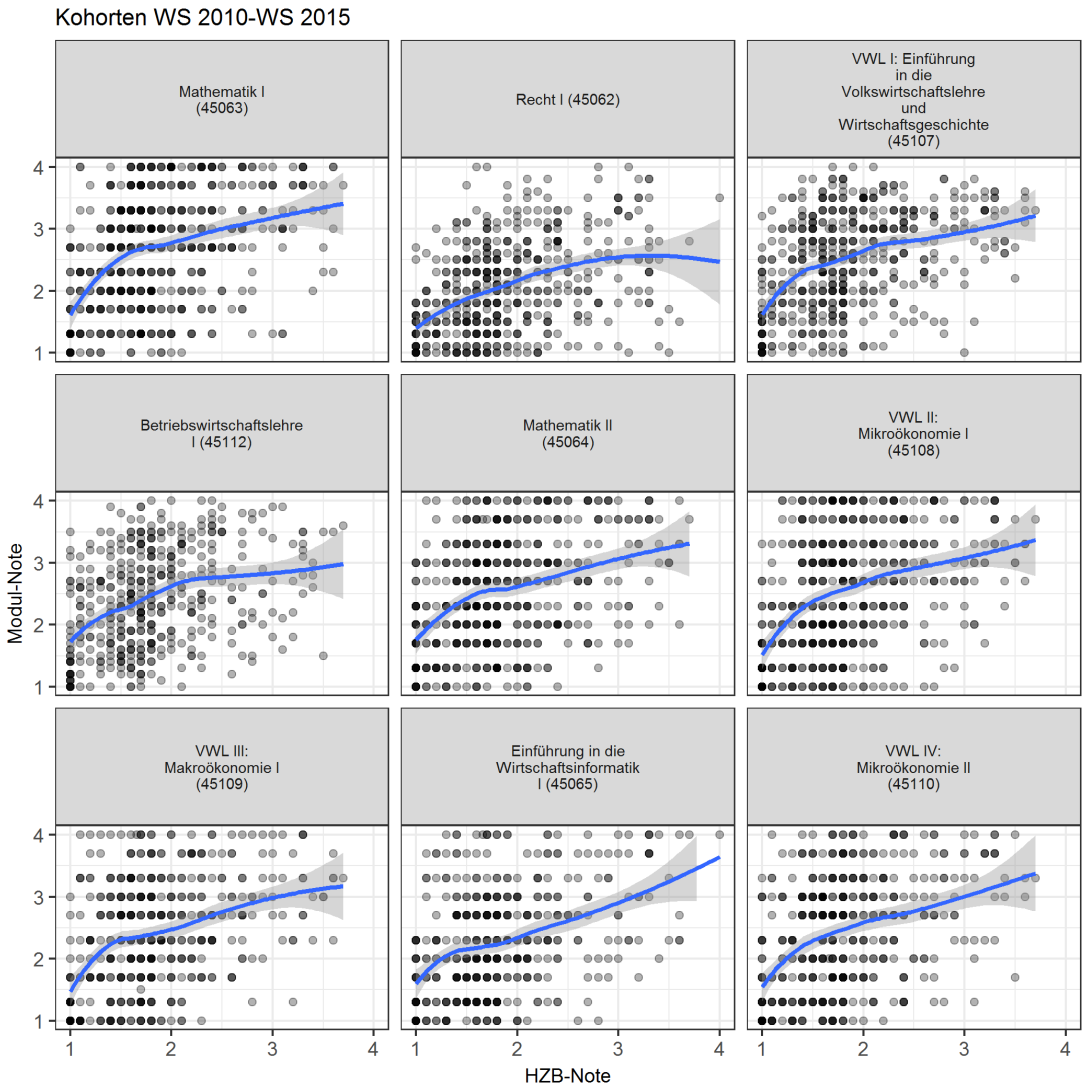


Abbildung 1.12: Erreichte Modulnoten nach Abiturnote (HZB-Note) in VWL. Scatterplot Smoother (LOESS) mit Konfidenzintervall

ebenfalls stark um die Abiturnote schwanken. Abbildung 1.13 zeigt auf Basis der Daten der HU[35] den Zusammenhang zwischen der Abiturnote (Note der Hochschulzulassung) und der erreichten Bachelor-Note. Für die BWL weist der Scatterplot Smoother keinen Zusammenhang von Schulnote und BA-Abschlussnote für einen weiten Bereich von Schulnoten aus. Wie bei der Entscheidung, ob der Studiengang vollendet wird oder abgebrochen wird, ist die Schulnote in dem Bereich bis zur Note 2.5 praktisch irrelevant. Dies ist aber der Bereich, wo die NC-Regelungen zwischen Zulassung und Nicht-Zulassung diskriminieren. Ein ganz anderes Bild ergibt sich bei der VWL: Hier ergibt sich ein deutlicher, fast linearer Zusammenhang zwischen der Abiturnote und der erreichten BA-Abschlussnote. Dies resultiert in einem R^2 von ungefähr 30%. Allerdings zeigt sich auch hier eine nicht unbeträchtliche Residualstreuung.

Auch Danilowicz-Gösele u. a. (2017) präsentieren in ihrer Figure 1 für Ökonomen Streudiagramme von Abiturnote und Abschlussnote mit einer hohen Residualstreuung. Ihr R^2 ist mit 0.247 sogar etwas geringer als der hier präsentierte Wert für den BA-VWL. Allerdings fasst die Analyse alle Abschlüsse am FB Wirtschaftswissenschaft zusammen und differenziert hierbei nicht zwischen Bachelor- und Masterabschlüssen und zwischen den Studiengängen BWL und VWL. Diese größere Heterogenität könnte für das geringere R^2 verantwortlich sein.

Insgesamt zeigt sich eine allenfalls schwache Prädiktion der Schulnote sowohl für den Abschluss des BA-Studiums als auch für die dabei erreichte BA-Note. Für den stark nachgefragten Studiengang BA-BWL mit strikten NC-Grenzen zeigt sich im NC-relevanten Bereich bis zur Note 2.5 überhaupt kein Zusammenhang von Schulnote und BA-Note. Für die VWL, wo der NC weitaus großzügiger ist, ist jedoch ein erkennbarer Zusammenhang von Schulnote und BA-Note gegeben, wenngleich die Residualstreuung immer noch so hoch ist, dass Prognosen auf den Studienerfolg wenig aussagekräftig sind.[36]

1.7 Resümee

Diese Analyse hat gezeigt, dass die Kombination von Befragungsdaten und administrativen Prüfungsdaten zahlreiche Ergebnisse zu Studienabbruchrisiken und deren Ursachen liefert, die mit retrospektiven Befragungsansätzen bzw. Querschnittbefragungen nicht oder nur um den Preis von hohen Nonresponsequoten und Messfehlern zu erhalten sind. Insbesondere zeigt sich, dass leistungsfähigere Studenten auch in Befragungen auskunftsbereiter sind. Dies ist ein Hinweis auf einen gewissen „Erfolgsbias“ bei Umfragen in der empirischen Bildungsforschung.

Weiterhin zeigt sich, dass die elterliche Bildung, mögliche Nebentätigkeiten während des Studiums sowie andere Hintergrundmerkmale keinen Einfluss auf den Studienabschluss haben.

[35] Wiederum wurden aus Datenschutzgründen aber auch zur Vergrößerung der Fallzahlen verschiedene Kohorten (WS2011/12 bis WS 2016/17) zusammengefasst.

[36] Über die Gründe für dieses unterschiedliche Studierverhalten kann hier nur spekuliert werden. Möglich wäre eine höhere Selektion der Studierenden in der VWL während ihres Studiums, so dass am Ende des Studiums als Resultat höherer Abbruchraten nur noch Studierende mit einer hohen fachlichen Motivation den BA-VWL abschließen.

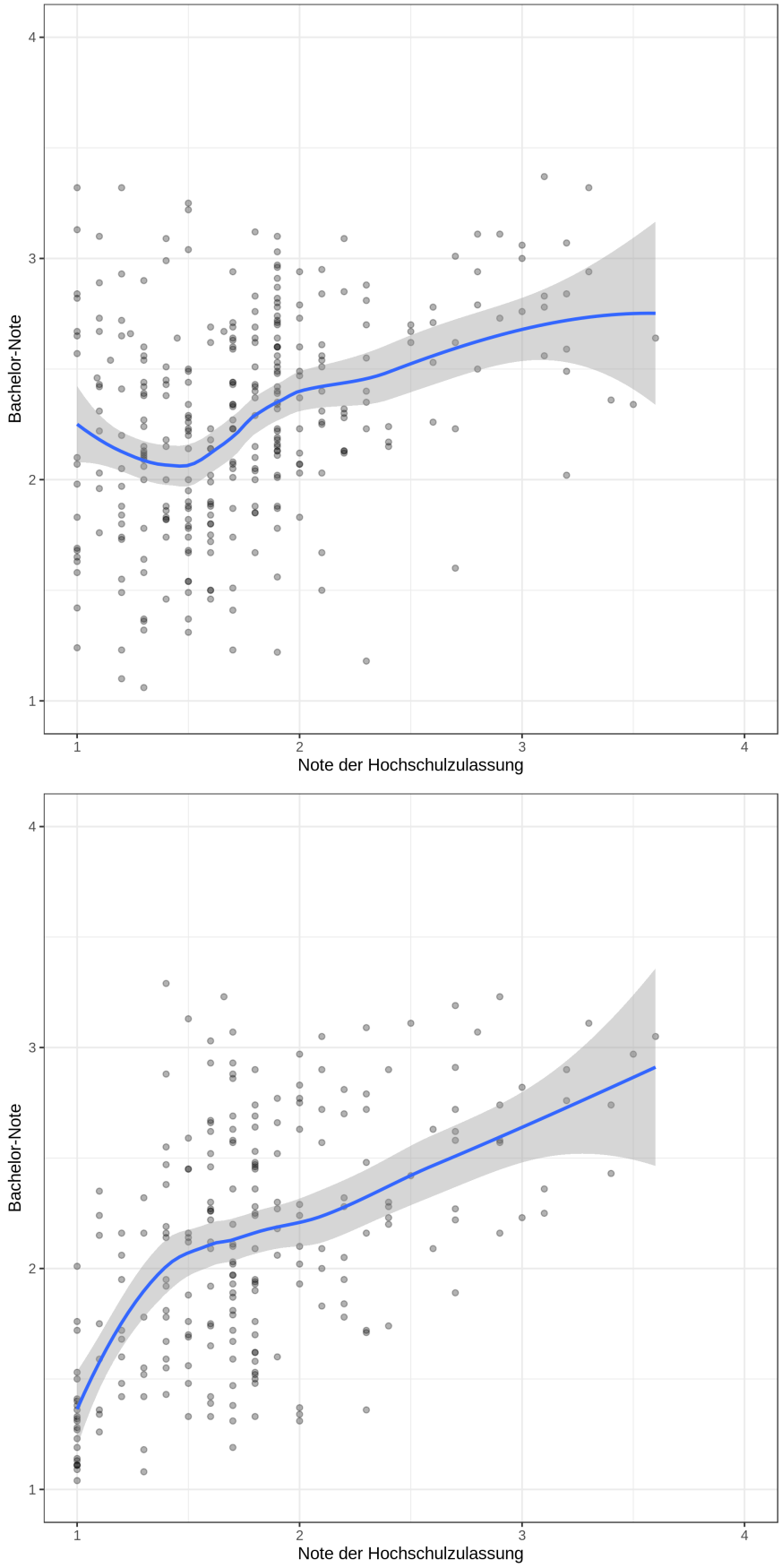


Abbildung 1.13: Der Einfluss der Abiturnote auf die BA-Note (HU Kohorten WS2011/12 bis WS 2016/17). Obere Grafik: BWL Untere Grafik: VWL

Dieser Befund steht in guter Übereinstimmung mit den Ergebnissen von Berens u. a. (2019), die ebenfalls keinen separaten Effekt von sozialen Proxy-Merkmalen (Krankenversicherung und lokale Kaufkraft) für den Studienerfolg finden. Die Tatsache, dass soziale Hintergrundmerkmale nur einen zu vernachlässigenden Einfluß auf den Studienerfolg haben, kann man als Indiz werten, dass die hier beteiligten Universitäten ihren Bildungsauftrag ohne soziale Hemmnisse erfüllen.

Allerdings sind die persönlichen Motivationsmerkmale (Sicherheit bei der Wahl des Studiengangs, Präferenz unter Alternativen und geäußerte Wechselabsicht) von hoher Relevanz für den Studienerfolg. Diese Merkmale sind zusammen mit den in der Studieneingangsphase erreichten Leistungspunkten ein sehr guter Indikator für den zukünftigen Studienerfolg.

Insgesamt zeigt sich, dass schon am Ende des ersten Studienjahres zuverlässige Vorhersagen über den Studienerfolg bis zum 8. Semester allein auf Basis der erzielten Leistungspunkte und einer persönlichen Einschätzung eines Studienwechsels möglich sind. Grundlage für diese hohe Prognosefähigkeit ist eine frühe, starke Bimodalität der Studierenden hinsichtlich der von ihnen erworbenen Studienpunkte. Diese Möglichkeit sollte aktiv für ein Mentorenprogramm genutzt werden. Es ist also nicht nötig, bis zum Ende der Regelstudienzeit zu warten, um gefährdete Studierende zu beraten. Eine umsetzbare Strategie würde die Studierenden, die eine gewisse Mindestanzahl von Leistungspunkten nicht erreicht haben, nach ihrer Studienmotivation fragen und sie mit ihren Prognosewerten für den Abschluss ihres Studiums konfrontieren.

Ein Wechsel des Studiengangs ist nicht per se negativ zu bewerten, obwohl der meist gebrauchte Begriff des „Studienabbruchs“ eine negative Konnotation besitzt. Häufig ist der Abbruch eines Studiengangs nur eine Station auf dem Weg zu einer endgültigen Berufsfindung.[37] Eine negative Bewertung als Vergeudung von finanziellen, aber auch persönlichen Ressourcen wie bei Berens u. a. (2019) ist nur dann angezeigt, wenn der Studienabbruch in einer späten Studienphase stattfindet. Aus diesem Grund ist eine Evaluation der Studierenden am Ende des ersten Studienjahres eine vielversprechende Möglichkeit, späte Studienabbrüche zu vermeiden.

Allerdings zeigen unsere deskriptiven Befunde, dass ein Großteil der Abbrüche relativ spät stattfindet. Teilweise wird bei einigen Studiengängen versucht, über Studien- bzw. Prüfungsordnungen späte Studienabbrüche zu vermeiden, indem man den Abschluss bestimmter Module innerhalb einer vorgegebener Frist vorschreibt.[38] Trotzdem ist eine derartige administrative Regelung, die erst in späteren Fachsemestern zum Tragen kommt, vermutlich langsamer als eine effiziente Evaluation nach dem ersten Studienjahr.

Der Einfluss der schulischen Bewertung über die Abiturnote[39] auf den Studienerfolg wird in der

[37] Dies ist der Bildungsgang eines der Autoren.

[38] Das Versäumen dieser Frist hat dann die gewollte Konsequenz, dass das Studium in diesem Studiengang beendet ist. Die hier berichteten Ergebnisse vom Fachbereich Wirtschaftswissenschaft der FU und der HU basieren auf Studienordnungen, die solche Fristenregelungen nicht kennen.

[39] Genauer die Hochschulberechtigungsnote

Literatur überwiegend akzeptiert, z.B. durch die Auswertung administrativer Prüfungsdaten, vgl. Danilowicz-Gösele u. a. (2017) oder als Metastudie von zahlreichen Auswertungen (Trapmann u. a. (2007). Diese Befunde beruhen jedoch fast immer auf einer Korrelationsanalyse, die die Linearität des Einflusses der Schulnote auf die Abschlussnote im Studium unterstellt.[40] Eine derartige Linearität ist aber in den von uns untersuchten Studiengängen nicht gegeben. Gerade in den Bereichen, wo der Numerus Clausus im Fach BWL harte Zulassungsgrenzen setzt, ist keine Differenzierung des Studienerfolgs über die Schulnote zu erkennen. D.h. Studienbewerber oberhalb und unterhalb der Zulassungsgrenze haben dieselben Aussichten auf einen Studienerfolg. Zudem gibt es bedeutende Unterschiede selbst für so verwandte Studiengänge wie BWL und VWL. Auch Danilowicz-Gösele u. a. (2017) konstatiert große Unterschiede zwischen einzelnen Fächergruppen. Unerwartet ist der Unterschied in Studiengängen in gemeinsamen Modulen.[41] Dies weist auf unterschiedliche fachliche Motivationen und Lernweisen in den Studiengängen hin, die den Studienerfolg beeinflussen.

Der Zugang zu den individuellen Prüfungsdaten ermöglicht auch die Darstellung über Streudiagramme von Abiturnoten und Prüfungsnoten. Diese weisen klar eine überragende Residualstreuung aus, nicht nur in den von uns analysierten Studiengängen sondern auch für andere Studiengänge, zum Beispiel in Figure 1 in Danilowicz-Gösele u. a. (2017). Als alleiniges Mittel zur Prognose des Studienerfolgs ist die Abiturnote mit einem R^2 von unter 0.20 wenig geeignet. Nimmt man die Studienmotivation und die Leistungsergebnisse der Studieneingangsphase mit in die Auswahl der Prognosemerkmale hinzu, so verschwindet der schulische Einfluss sofort.

Man fragt sich daher, was der Abiturnote so eine überragende Bedeutung bei der Zulassung zum Studium in Deutschland gibt. Nach Danilowicz-Gösele u. a. (2017) gibt es einfach keine besseren Prädiktoren vor Studienbeginn. Allerdings wurde die alleinige Anwendung der Schulnote vom Bundesverfassungsgericht als Kriterium bei der Zulassung zum Medizinstudium als verfassungswidrig verworfen.[42] Unsere Ergebnisse belegen einen hohen prognostischen Wert der Studienergebnisse des ersten Semesters und der Selbsteinschätzung der Studierenden. Diese Informationen könnte man für eine probeweise Zulassung nutzen, die nach dem ersten Studienjahr evaluiert wird. Eine Art Probestudium[43] gibt es bereits an der FU (siehe `https://www.fu-berlin.de/universitaet/kooperationen/schulen/studierende/index.htmlaufgerufenam15.7.2020`. In Frankreich aber auch anderen europäischen Ländern ist eine Evaluation der Studierenden nach dem ersten Studienjahr die Regel. Die technologischen Möglichkeiten für eine Vergrößerung der Vorlesungszahlen sind mit der Einführung der Online-Lehre und elektronischen Prüfungsräumen im Gefolge der Corona-Krise deutlich gewachsen. Frühere Restriktionen durch die Größe der Hörsäle können damit überwunden werden. Insgesamt ist damit der Spielraum für alternative Zulassungsverfahren gestiegen.

[40] Es werden fast ausschließlich Korrelationskoeffizienten berichtet.

[41] So der unterschiedliche Einfluss der Abiturnote auf die Note im Modul Recht für die BWl und die VWL.

[42] vgl. `https://www.tagesschau.de/inland/medizinstudium-verfassungsgericht-101.html` aufgerufen am 3.12.2019

[43] Manchmal wird auch die Bezeichnung „Schnupperstudium“ gewählt.

1.8 Appendix

1.8.1 Verwendete Variablen

Tabelle 1.9: Übersicht über die zur Selektion verfügbaren Variablen, für die Modelle aus den Tabellen 1.3, 1.5, 1.6, 1.8.

Variable	Fragenummer	wechsel Ordinal (1.5)	wechsel Binär(1.6)	Abbruch (1.8)	Leistungspunkte(1.3)
Studiengang	42	✓	✓	✓	✓
Sem_1_LP	26	✓	✓	✓	
lehre	3	✓	✓	✓	✓
fh	4	✓	✓	✓	✓
abi_ort	5	✓	✓	✓	✓
bafg	8	✓	✓	✓	✓
nebenerw	9	✓	✓	✓	✓
wstd	5	✓	✓	✓	✓
stud_rank	11	✓	✓	✓	✓
stud_sich	13	✓	✓	✓	✓
uni_rank	14	✓	✓	✓	✓
stund_inf_osa	15	✓	✓	✓	✓
stund_inf_sonst	15				✓
studmot_ver	21	✓	✓	✓	
studvorst_ver	22	✓	✓	✓	✓
stud_wechsel	23			✓	
alternative	24			✓	
geschl	41	✓	✓	✓	✓
HZB-Note	1			✓	✓
HZB-Note_kat	1	✓	✓	✓	✓
abschluss_eltern	6+7	✓	✓	✓	✓
alter	40			✓	✓
alter_kat	40	✓	✓	✓	✓
wechselneigung	23			✓	
schwerpunkt	16	✓	✓	✓	

1.8.2 Dokumentation Fragebogen

Freie Universität Berlin

fu:stat

Fragebogen Nr.: 1

Hörsaalbefragung zur Verbesserung der Studiensituation im Fachbereich Wirtschaftswissenschaften

Schulischer Hintergrund

1. Abitur bzw. Hochschulzulassungsnote? ☐,☐☐

2. Leistungsfächer?

_____

_____

3. Lehre Absolviert Ja ☐ Nein ☐

4. FH Abschluss Ja ☐ Nein ☐

5. Ort des Erwerbs der Hochschulreife

Berlin	andere Bundesländer	EU	nicht EU
☐	☐	☐	☐

Sozialer Hintergrund

6. Höchster Bildungsabschluss Vater

Studium	Abitur	Mittlere Reife oder weniger
☐	☐	☐

Abbildung 1.14: Dokumentation des Fragebogens: Seite 1

Fragebogen Nr: 1

7. Höchster Bildungsabschluss Mutter

Studium	Abitur	Mittlere Reife oder weniger
□	□	□

Studienfinanzierung

8. Bafög Ja □ Nein □

9. Nebenerwerb Ja □ Nein □

10. Falls ja: Umfang in Std. pro Woche □□

Studienmotivation

11. Die wievielte Wahl war Ihr gewählter Studiengang?

1	2	3
□	□	□

12. Falls Ihr gewählter Studiengang nicht die erste Präferenz war, welcher war dann die erste Präferenz?

_____

13. Wie sicher waren Sie sich bei der Wahl Ihres Studienganges?

	5	4	3	2	1	
	□	□	□	□	□	
sehr **sicher**						sehr **unsicher**

14. Die wievielte Wahl war die FU als Uni bei Ihnen?

1	2	3
□	□	□

15. Welche Möglichkeiten, sich über Ihren Studiengang zu informieren, haben Sie genutzt?

Online Studienassistent (OSA)	□
sonstige Quellen	□

Welche? _____

_____

16. Haben Sie Ihr Studium aufgrund bestimmter Module oder möglicher Schwerpunkte an der FU gewählt (Informationen aus Studiengangsbeschreibungen, Studienverlaufsplänen, etc.)?

ja □ nein □

2

Abbildung 1.15: Dokumentation des Fragebogens: Seite 2

17. Falls ja, welche Module / Schwerpunkte interessierten Sie besonders?

Management	☐
Marketing	☐
FACTs	☐
Statistik / Ökonometrie	☐
Jura	☐
Mikroökonomie (Monopole, Oligopole, ...	☐
Makroökonomie (Zentralbanken, Arbeitslosigkeit, ...)	☐
Wirtschaftstheorie / -geschichte	☐

18. Werden Module in Ihrem Interessensbereich aktuell ausreichend angeboten? Ja ☐ Nein ☐

19. In welchen Themengebieten würden Sie sich ein Angebot wünschen, dass über die aktuellen Wahlmöglichkeiten hinausgeht?

Management	☐
Marketing	☐
FACTs	☐
Statistik / Ökonometrie	☐
Jura	☐
Mikroökonomie (Monopole, Oligopole, ...	☐
Makroökonomie (Zentralbanken, Arbeitslosigkeit, ...)	☐
Wirtschaftstheorie / -geschichte	☐
Anderer Bereich? _____	
Ich bin zufrieden.	☐

20. Bitte ordnen Sie die folgenden Fachschwerpunkte nach ihrer Relevanz für Ihren Studiengang (niedrig 1-8 hoch)

Merkmal	**Wichtigkeit**
Management	[ ]
Marketing	[ ]
FACTs	[ ]
Statistik / Ökonometrie	[ ]
Jura	[ ]
Mikroökonomie	[ ]
Makroökonomie	[ ]
Wirtschaftstheorie / -geschichte	[ ]

21. Inwiefern hat sich Ihre Studienmotivation zum Beginn des zweiten Semesters geändert?

starke Zunahme		keine Veränderung		starke Abnahme
☐	☐	☐	☐	☐

22. Inwiefern haben sich Ihre Vorstellungen bezüglich Ihres Studienganges bestätigt?

positiv Überrascht	positiv Bestätigt	negativ Bestätigt	negativ Überrascht
☐	☐	☐	☐

Abbildung 1.16: Dokumentation des Fragebogens: Seite 3

Fragebogen Nr: 1

23. Bedenken Sie einen Wechsel des Studienganges?

ja	unsicher	nein
☐	☐	☐

24. Falls Sie einen Wechsel erwägen, welchen Weg würden Sie alternativ einschlagen?

Berufseinstieg	☐
Rückkehr in den Beruf	☐
Anderer Studiengang	☐
Welcher? _____	

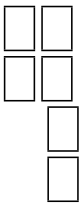
Studienerfolg im ersten Semester

25. Geplante Leistungspunkte

26. Erreichte Leistungspunkte

27. Anzahl nicht bestandener Prüfungen, ohne geplantes Nichtbestehen

28. Anzahl der geplant nicht bestandene Prüfungen



Kritikpunkte am Studium

Studienorganisation

29. Prüfungszeitraum

30. Lehrveranstaltungen

31. Tutorien

Studieninhalte

32. unbeliebte Veranstaltungen

Abbildung 1.17: Dokumentation des Fragebogens: Seite 4

33. Art des Lernens

34. Praxisbezug

35. Flexibilität

36. Sonstiges

Wenn Sie drei Wünsche frei hätten...

37. Erster Wunsch

38. Zweiter Wunsch

39. Dritter Wunsch

Zu Ihrer Person

40. Wann wurden Sie geboren? □□

41. Geschlecht weiblich □ männlich □

42. Studiengang BWL □ VWL □ sonstiges □

43. Fachsemester □□

44. Hochschulsemester □□

Abbildung 1.18: Dokumentation des Fragebogens: Seite 5

Part II

Combining Data for Prediction

Chapter 2

Data-driven Transformations in Small Area Estimation

2.1 Introduction

Model-based methods for small area estimation (SAE) are now widely used in practice for producing reliable estimates of linear and non-linear indicators for areas/domains with small sample sizes. Examples of indicators that are estimated by using model-based methods include poverty (income deprivation) and inequality measures such as the head count ratio, the poverty gap and the income quintile share ratio. Two popular small area methods in this case are the empirical best predictor (EBP), proposed by Molina and Rao (2010) and the World Bank method, proposed by Elbers, Lanjouw, et al. (2003). Both approaches are based on the use of unit-level linear mixed regression models. Although estimation of complex indicators can be also implemented with area-level linear mixed regression models (Fabrizi and Trivisano, 2016; Schmid, Bruckschen, et al., 2017), in this paper we focus on unit-level linear mixed regression models. In the original paper, Molina and Rao (2010) assumed that the error terms of the linear mixed regression model follow a Gaussian distribution. In case the model error terms significantly deviate from normality, the EBP estimator can be biased. What are the options available to the data analyst when the normality assumptions are not met? One option is to formulate the EBP under alternative and more flexible parametric assumptions. Graf et al. (2019) study an EBP method under the generalized beta distribution of the second kind (GB2), whereas Diallo and Rao (2014) propose the use of skewed-normal distributions in applications with income data. One complication with using the EBP under alternative parametric distributions is that new tools for estimation must be developed and training for the data analyst is needed. In addition, misspecification of the model assumptions is still possible. Another option when the Gaussian assumptions are not satisfied is to use a methodology that minimizes the use of parametric assumptions. For instance, Elbers and Weide, 2014 proposed an EBP method based on normal mixture models. With this method the distribution of the error terms is described by normal mixtures. Weidenhammer et al. (2014) recently proposed a method that aims at estimating the

quantiles of the empirical distribution function of the data. The estimation of the quantiles is facilitated by a nested error regression model using the asymmetric Laplace distribution for the unit-level error terms as a working assumption. The estimation of the random effects can be made completely non-parametric by using a discrete mixture proposed by Marino, Tzavidis, et al. (2018) and Marino, Ranalli, et al. (2019). Another option, and the one we study in this paper, is to find an appropriate transformation such that the model assumptions (in this paper the Gaussian assumptions of the EBP method) hold. The aim is to find transformations that (a) are data-driven and optimal according to some criterion and (b) can be implemented by using standard software. To the best of our knowledge, the use and choice of transformations in SAE has not been extensively studied or it has been studied in fairly ad-hoc manner. Elbers, Lanjouw, et al. (2003) and Molina and Rao (2010) suggested the use of logarithmic-type transformations for income data. However, are such transformations the most appropriate choice? Can alternative transformations offer improved estimation? In order to answer these research questions, the paper investigates data-driven transformations for small area estimation.

The choice of transformations when modelling income-type outcomes presents different challenges. Transformations should be suitable for dealing with unimodal, leptokurtic and positively skewed data that may include zero and negative values. Besides the logarithmic transformation and its modifications (e.g. the log-shift transformation) a popular family of data-driven transformations that includes the logarithmic one as a special case is the Box-Cox family (Box and Cox, 1964). Since the Box-Cox transformation is not defined for negative values, when negative values are present, the data must be shifted to the positive range. Another difficulty with the use of the Box-Cox transformation is the truncation on the transformation parameter described later in Section 2.4. A solution to this problem can be offered by using of the dual power transformation. Although extensive literature on the use of transformations exists, see for example, John and Draper (1980), Bickel and Doksum (1981) and Yeo and Johnson (2000) among others. In this paper we focus on three types of transformations, namely log-shift, Box-Cox and dual power transformations.

In addition to selecting the type of transformation, estimating the transformation parameter adds another layer of complexity. To the best of our knowledge the use of transformations in recent applications of SAE has employed visual residual diagnostics for finding a suitable transformation parameter. In this paper we propose a structured, data-driven approach for estimating the transformation parameter. In particular, we introduce maximum likelihood and residual maximum likelihood methods for estimating the transformation parameter under the linear mixed regression model following Gurka et al., 2006. Alternative estimation approaches based on the minimization of distances (Cramér, 1928; Chakravarti and Laha, 1967) and on the minimization of the skewness (Carroll and Ruppert, 1987) are also discussed.

At this point we should emphasise some of the differences between the present paper and the paper published by Tzavidis, Zhang, et al., 2018. The latter paper proposes a general framework for the production of small area statistics including measuring uncertainty. Broadly

speaking, the proposed framework is based on three stages, namely specification of the problem, analysis of the data and adaptation of the model, and method evaluation. The paper focuses on practical aspects of the SAE process and not on proposing new methodology. The target audience includes practitioners using small area estimation methods for example, colleagues in National Statistical Institutes. The use of transformations, as a parsimonious approach to adapting the model, is mentioned in the paper but the methodological details are not derived. In contrast, the present paper focuses on developing new and generally applicable methodology that underpins the use of data-driven transformations in SAE and applies the methodology to real data problems. In particular, the current paper proposes the use of the empirical best predictor (EBP) (Molina and Rao, 2010) with data-driven transformations estimated with likelihood-based methods. The paper focuses on scaled transformations that allow the use of standard software for small area estimation. As we mentioned above, the focus is on the use of the log-shift, Box-Cox, and dual power transformations and the mathematical derivations for developing scaled transformations are presented. Illustrating how to derive scaled transformations for these three transformation types will allow researchers to use similar developments for other families of transformations. In addition, in the present paper we propose two bootstrap schemes (parametric- and wild-type) for estimating the MSE under data-driven transformations and extend these to capture the additional uncertainty due to the estimation of the transformation parameter. The wild bootstrap scheme can been viewed as an insurance policy in case there are some "mild" departures from Normality after using transformations. Finally, the present paper includes results from model-based simulation studies that are necessary for comparing the performance of data-driven transformations against the use of fixed, ad-hoc transformations. Emphasis is given to the estimation of poverty and inequality indicators due to their important socio-economic relevance and policy impact. We further study whether the impact of departures from Gaussian assumptions is different depending on the target of estimation. For instance, departures from normality may have lesser impact on estimates of median income compared to estimates indicators that are more sensitive in the data distribution. The use of model-based simulations was one of the method evaluation approaches recommended in the paper by Tzavidis, Zhang, et al., 2018.

The rest of the paper is structured as follows. The EBP approach is introduced in Section 2.2. Section 2.3 presents the survey data we use in this paper and makes the case, via the use of residual diagnostics, for using transformations. In Section 2.4 selected transformations are introduced and extended for their use with model-based SAE methods under the linear mixed regression model. This section includes the theoretical details about the choice of an appropriate scale and estimation of the transformation parameter. MSE estimation is discussed in Section 2.5. In Section 2.6 the proposed methods are applied to data from Guerrero in Mexico for estimating a range of deprivation and inequality indicators and corresponding estimates of uncertainty. In Section 2.7 the proposed methods are further evaluated by realistic - for income data - model-based simulations. Section 2.8 summarizes the main findings and outlines further research.

2.2 The Empirical Best Prediction (EBP) method

Let U denote a finite population of size N partitioned into D areas or domains (representing the small areas) $U_1, U_2, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, where $i = 1, \ldots, D$ refers to the ith area. Let y_{ij} be the target variable defined for the jth individual belonging to the ith area, with $j = 1, \ldots, N_i$. Denote by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ the design matrix containing p explanatory variables and define by s as the set of sample units, with s_i the in-sample units in area i and by r be the set of non-sampled units, with r_i the out-of-sample units in area i. Let n_i denote the sample size in area i with $n = \sum_{i=1}^{D} n_i$. Hence, we define by $\mathbf{y}_i$ a vector with population elements of the target outcome for area i partitioned as $\mathbf{y}_i^T = \left(\mathbf{y}_{is}^T, \mathbf{y}_{ir}^T\right)$, where $\mathbf{y}_{is}$ and $\mathbf{y}_{ir}$ denote the sample elements s and the out-of-sample elements r in area i respectively. Let us now describe in more detail the EBP approach by Molina and Rao (2010), which is the methodology we focus on in this paper. Under this approach census predictions of the target outcome are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. The point of departure is the standard parametric unit-level linear mixed regression model, which is also known as the unit-level nested error regression model. This is defined by Battese et al. (1988) as:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2), \tag{2.1}$$

where u_i, the area-specific random effects, and e_{ij}, the unit-level error, are assumed to be independent. Assuming normality for the unit-level error and the area-specific random effects, the conditional distribution of the out-of-sample data given the sample data is also normal. A Monte Carlo approach is used to obtain a numerically efficient approximation to the expected value of this conditional distribution as follows:

1. Use the sample data to obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and the weighting factors $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

2. For $l = 1, \ldots, L$:

   (a) Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1 - \hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and obtain a pseudo-population of the target variable by:

   $$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)},$$

   where the predicted random effect $\hat{u}_i$ is defined as $\hat{u}_i = E(u_i | \mathbf{y}_{is})$.

   (b) Calculate the indicator of interest $I_i^{(l)}$ in each area.

3. Finally, take the mean over the L Monte Carlo runs in each area to obtain a point estimate of the indicator of interest:

   $$\hat{I}_i^{EBP} = \frac{1}{L} \sum_{l=1}^{L} I_i^{(l)}.$$

As is common in real applications, some areas are out-of-sample. For those areas, we cannot estimate an area-specific random effect, and hence the corresponding area-specific random effect is set equal to zero. Synthetic values of the outcome for the out-of-sample areas are then generated under the linear mixed regression model as follows:

$$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(l)} + e_{ij}^{(l)},$$

with $u_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$. Finally, a parametric bootstrap - under the assumed model - is used for the MSE estimation. This is discussed in some detail in Section 2.5. Assuming normality for the error terms is a convenient assumption as allows the conditional distribution of $\mathbf{y}_r|\mathbf{y}_s$ to be derived. However, in applications that involve modelling an income-type outcome, as is the case in this paper, assuming normality is unrealistic. If our primary target is to develop a methodology that can easily be used in practice, finding appropriate data transformations is important.

2.3 The Guerrero case study: Data source and initial analysis

In this section, we describe the data sources used in the application and provide a motivation for the use of transformations. The case study was carried out by using the open-source software R (R Core Team, 2017) and R packages. The data we use in this paper come from the Mexican state of Guerrero, one of the 32 states in Mexico. The state Guerrero is considered by the World Bank to be one of the states - next to the State of Mexico investigated by Tzavidis, Zhang, et al., 2018 - mostly contributing to income inequality in Mexico (Bedoya et al., 2013). Additionally, according to the United Nations Development Programme (UNDP), Guerrero has one of the highest rates of poverty and lack of infrastructural development (Tortajada, 2006). According to the general social development law in Mexico, the National Institute of Statistics and Geography (INEGI) has to provide relevant official statistics at the national, state and municipal-levels. Furthermore, the Social Development Law (SDL) in Mexico establishes that the National Council for the Evaluation of Social Development Policy (CONEVAL) should measure poverty at state level every two years and at municipal level every five years. For carrying out the analysis the statistical and geographical information was provided by INEGI through the Household Income and Expenditure Survey (ENIGH) 2010 and the National Population and Housing Census of 2010. Looking in more detail at the data available and their geographic coverage, Guerrero comprises 81 administrative divisions, known as municipalities. From the 81 municipalities 40 municipalities with 1611 households are in-sample (in the sample of the ENIGH survey) and the remaining 41 municipalities are out-of-sample. For the in-sample municipalities the maximum sample size in a municipality is 511, the minimum is 9 and the median is 24 households. Note that more than 30% of the sample is from a single municipality, the capital (Chilpancingo de los Bravo).

The survey and census data include a large number of socio-demographic variables, which are common and are measured similarly in both data sources. The total household per capita in-

come (*ictpc*, measured in pesos) is a variable recorded for households and is available in the survey but not in the census. We used this variable as a proxy that best approximates the living standard in Guerrero and as the outcome variable in our models. Socio-economic variables available for the households both in the survey and census data are used as explanatory variables. The underlying linear mixed regression model (2.1) of the EBP has two levels, households and municipalities. The variables available in the survey and census data, which are identified by using Bayesian information criterion (BIC) as good predictors of *ictpc*, are described in Table 2.1. From now on, the working model is assumed to be known and fixed.

Table 2.1: Description of the explanatory variables used in the working model

Determinant	Variable
Occupation	1) Indicator if the head of household and the spouse are employed
	2) Type of household occupation
	3) Total number of employees older than 14 years in a household
	4) Percentage of employees older than 14 years in a household
Sources of income	5) Indicator of a household receiving remittances
Socioeconomic level	6) Availability of assets in the household
	7) Total number of goods in the household
Education	8) Average standardized years of schooling (by age and sex) within the household relative to the population

The next step after the identification of a possible set of covariates is assessing the predictive power of the model. Nakagawa and Schielzeth (2013) propose the use of two coefficients of determination suitable for linear mixed regression models: (a) the marginal R^2_m, which is a measure for the variance explained by fixed effects and (b) the conditional R^2_c, which measures the variance explained by both, the fixed and random effects. Without using any transformation, these measures are both around 35% and the corresponding intraclass correlation (ICC) under the model is 0.027.

In order to explore the validity of the Gaussian assumptions underlying the linear mixed regression model, it is common practice to perform normality tests and some residual diagnostics. The p-values of the Shapiro-Wilk (S-W) test statistic are equal to $2.2 \cdot 10^{-16}$ for the household-level and 0.197 for the municipal-level. These results indicate that the null hypothesis of normality for the household-level is rejected. As normality tests like Shapiro-Wilk have some problems we also present some visual approaches in addition. Figure 2.1 presents the Normal probability quantile-quantile (Q-Q) plots for household-level and municipal-level residuals. As expected, in the case of using the non-transformed *ictpc* variable, the shape of the Q-Q plots is clearly different from what would be expected under normality. In addition, the analysis of skewness and kurtosis for both error terms is also informative. The skewness and kurtosis for a Normal distribution are equal to zero and three, respectively. The skewness and kurtosis of the household-level are equal to 6.338 and 75.483, and for the municipal-level equal to 0.448 and 3.250. These results indicate severe departures - especially for the household-level - from the Gaussian assumptions when modelling the non-transformed income.

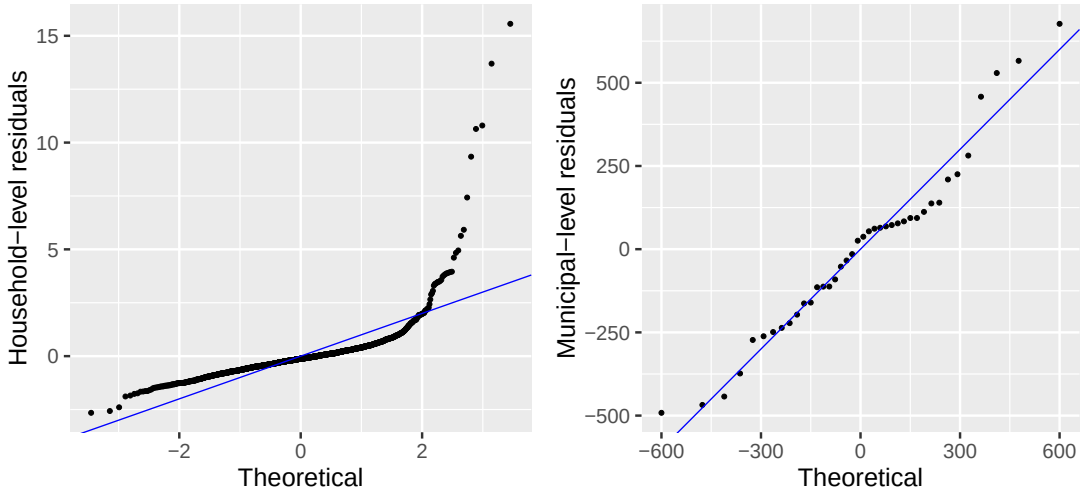


Figure 2.1: Q-Q plots of the household- and municipal-level error terms

2.4 Use of transformations

In order to get closer to normality, it is common to use a one-to-one transformation $T(y_{ij}) = y^*_{ij}$ of the target variable. The application of the natural logarithmic transformation, which is a popular choice for income data, leads in many cases from right-skewed to more symmetric distributions. This approach is followed by the paper by Molina and Martín, 2018, in which the logarithmic transformation is applied to an income-type variable for meeting the assumptions of the model proposed by Battese et al., 1988. In particular, the paper by Molina and Martín, 2018 proposes analytic MSE estimators and develops bias correction terms necessary when using estimating small area averages using a logarithmic transformation under the linear mixed regression model. The logarithmic transformation is frequently used for dealing with non-normality due to its simplicity. However, can an alternative transformation with data-driven parameter(s) λ, $T_\lambda(y_{ij}) = y^*_{ij}(\lambda)$, offer small area estimates with improved precision?
The structure of the section is as follows. In Section 2.4.1 we introduce the EBP approach with data-driven transformations. In Section 2.4.2 we propose likelihood-based approaches for estimating the transformation parameter, λ, in general and discuss three particular subcases - the log-shift, Box-Cox and dual power transformations - in detail. Finally, in Section 2.4.3 we discuss alternative to likelihood-based approaches for estimating the transformation parameter.

2.4.1 EBP under transformations

In order to apply the EBP method by using transformations, the linear mixed regression model is re-defined as follows:

$$y^*_{ij}(\lambda) = \mathbf{x}^T_{ij}\boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N\big(0, \sigma^2_u\big) \text{ and } e_{ij} \overset{iid}{\sim} N\big(0, \sigma^2_e\big). \tag{2.2}$$

The EBP approach under transformations can be re-written as follows:

1. Select a transformation and obtain $T_\lambda(y_{ij}) = y^*_{ij}(\lambda)$.
2. Use the transformed sample data to obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_u, \hat{\sigma}^2_e$ and calculate the weighting factors, $\hat{\gamma}_i = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_u + \frac{\hat{\sigma}^2_e}{n_i}}$.
3. For $l = 1, \ldots, L$:
   (a) Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}^2_u(1-\hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}^2_e\big)$ and obtain a pseudo-population of the target variable by:
   $$y_{ij}{}^{*(l)} = \mathbf{x}^T_{ij}\hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)}.$$
   (b) Back-transform $y_{ij}^{*(l)}$ to the original scale $y_{ij}^{(l)} = T_\lambda^{-1}\left(y_{ij}^{*(l)}\right)$.
   (c) Calculate the indicator of interest $I_i^{(l)}$ in each area.
4. Finally, take the mean over the L Monte Carlo generations in each area to obtain an estimate of the indicator of interest:
   $$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L} I_i^{(l)}.$$

2.4.2 Likelihood-based approach for estimating λ

For estimating the transformation parameter λ, the linear mixed regression model defined in (2.2) is used. Assume that the transformed vectors $\boldsymbol{y}^*_i$ are independent and normally distributed for some unknown λ,

$$\boldsymbol{y}^*_i(\lambda) \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i) \quad \text{for} \quad i = 1, \ldots, D,$$

where

$$\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}_i = \sigma^2_u \mathbf{1}_{N_i}\mathbf{1}'_{N_i} + \sigma^2_e \mathbf{I}_{N_i},$$

with $\mathbf{1}_{N_i}$ a column vector of ones of size N_i and $\boldsymbol{I}_{N_i}$ the $N_i \times N_i$ identity matrix, the vector of unknown model parameters is $\boldsymbol{\theta}^T = (\boldsymbol{\beta}, \sigma^2_u, \sigma^2_e, \lambda)$. The log-likelihood function under the model is defined as follows:

$$\begin{aligned} l_{\text{ML}}(\boldsymbol{y}^*, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}^*_i(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}^*_i(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]. \end{aligned}$$

The log-likelihood function in relation to the original observations is obtained by multiplying the normal density by the log of the Jacobian of the transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}^*_i(\lambda)$. The

Jacobian $J(\lambda, \boldsymbol{y})$ is defined as $\prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right|$ and is incorporated as follows:

$$\begin{aligned} l_{\text{ML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] + \log J(\lambda, \boldsymbol{y}). \end{aligned}$$

The maximization of $l_{\text{ML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta})$ produces maximum likelihood (ML) estimates of the unknown parameters $\boldsymbol{\theta}$. However, in the theory of linear mixed regression models, when interest focuses on accurate estimators of the variance components, restricted maximum likelihood (REML) theory is recommended (Verbeke and Molenberghs, 2000). The REML is defined as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] + \log J(\lambda, \boldsymbol{y}). \quad (2.3) \end{aligned}$$

The use of the scaled version of a selected transformation, defined by $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{\frac{1}{n}}} = z_{ij}^*(\lambda)$, is crucial for estimating the transformation parameter λ under the REML approach presented above. The Jacobian of such a scaled transformation is equal to 1. This means, the scale of the likelihood is preserved independently of the transformation and its parameter λ. Therefore, values of the log-likelihood function - under differently transformed $y_{ij}^*(\lambda)$ - can be directly compared and the log-likelihood function simplifies to the log-likelihood function of the linear mixed regression model. As a result, standard software for fitting this model can be used to estimate the transformation parameter λ. Eventhough using scaled transformations aids the implementation of the methods in practice, appropriate scaling factors must be developed depending on the type of transformation used.

Although the theory is applicable to data-driven transformations in general, we focus on the three types of transformations we presented in Section 2.1, namely the log-shift, Box-Cox and dual power transformations. Additionally, we use the frequently applied logarithmic transformation as a benchmark. This transformation is defined by $y_{ij}^* = \log(y_{ij} + s)$, where s denotes a fixed parameter such that $y_{ij} + s > 0$. The log-shift transformation (Yang, 1995), presented below, extends the logarithmic transformation by including the data-driven transformation parameter, $\lambda \geq s$, which needs to be estimated:

$$y_{ij}^*(\lambda) = \log(y_{ij} + \lambda).$$

When $\lambda = s$, the logarithmic transformation is obtained. The Box-Cox transformation (Box and Cox, 1964) is defined as follows:

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

where s denotes a fixed parameter such that $y_{ij} + s > 0$. If $\lambda = 0$, the logarithmic transformation is then a special case and if $\lambda = 1$, the data are only shifted. One difficulty with the Box-Cox type transformations is the long-standing truncation, i.e. $y_{ij}^*(\lambda)$ is bounded, from below by $\frac{1}{\lambda}$ if $\lambda > 0$ and from above by $\frac{-1}{\lambda}$ if $\lambda < 0$. This is the key motivation for the third type of transformation. The dual power transformation, introduced by Yang (2006), is defined as follows:

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0, \\ \log(y_{ij} + s) & \text{if } \lambda = 0, \end{cases}$$

where s is defined as in the case of the Box-Cox transformation.

The corresponding Jacobian used in (2.3) and scaled versions of the log-shift, Box-Cox and dual power transformations are presented in Table 2.2. For more details we refer to the developments in Appendix 2.9.1.

Table 2.2: Jacobian and scaled data-driven transformations for log-shift, Box-Cox and dual

Transformation	Jacobian J	Scaled transformation $z_{ij}^*(\lambda)$
Log-Shift	$\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+\lambda)^{-1}$	$J^{\frac{-1}{n}}\log(y_{ij}+\lambda)$
Box-Cox	$\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}^{\lambda-1}$	$J^{\frac{-1}{n}}\frac{(y_{ij}+s)^\lambda-1}{\lambda}$, if $\lambda \neq 0$ $J^{\frac{-1}{n}}\log(y_{ij}+s)$ if $\lambda = 0$
Dual	$\frac{\prod_{i=1}^{D}\prod_{j=1}^{n_i}\left((y_{ij}+s)^{\lambda-1}+(y_{ij}+s)^{-\lambda-1}\right)}{2}$	$J^{\frac{-1}{n}}\frac{(y_{ij}+s)^\lambda-(y_{ij}+s)^{-\lambda}}{2\lambda}$, if $\lambda \neq 0$ $J^{\frac{-1}{n}}\log(y_{ij}+s)$ if $\lambda = 0$

2.4.3 Alternative approaches for estimating λ

The ML and REML approaches introduced in 2.4.2 rely on parametric assumptions that may be influenced by outliers in the data. The kurtosis and skewness are crucial features for defining the shape of a distribution and a proximity measure can be minimized in order to find a transformation parameter under which the empirical distribution of residuals has skewness and kurtosis as close as possible to zero and three respectively. In general, skewness is considered more important than kurtosis, therefore, minimizing the skewness is an approach already

considered in the literature (Royston and Lambert, 2011) for linear models as follows:

$$\hat{\lambda}_{\text{skew}} = \underset{\lambda}{\operatorname{argmin}} \, |S_{e_\lambda}|,$$

where S_{e_λ} is the skewness and $\sigma^2_{e_\lambda}$ denotes the variance of the unit-level error terms. Note that the index λ is used to emphasize that the skewness and the variance parameters depend on the transformation parameter. In the context of linear mixed regression models, an additional problem arises as there are two independent error terms to be considered. We propose a pooled skewness approach that uses a weight w to ensure that the larger the error term variance $\sigma^2_{e_\lambda}$ is, the more weight its skewness will have in the minimization. Let S_{u_λ} be the skewness and $\sigma^2_{u_\lambda}$ be the variance of the area-specific random effects u_i of the linear mixed regression model. The estimation criteria in the pooled skewness approach is defined as follows:

$$\hat{\lambda}_{\text{poolskew}} = \underset{\lambda}{\operatorname{argmin}} \, \Big(w|S_{e_\lambda}| + (1-w)|S_{u_\lambda}| \Big),$$
$$\text{where } w = \frac{\hat{\sigma}^2_{e_\lambda}}{\hat{\sigma}^2_{u_\lambda} + \hat{\sigma}^2_{e_\lambda}}.$$

Considering only the skewness may ignore other properties of the distribution. Hence, a measure describing the distance between two distribution functions is another alternative. Two distance measures, the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) are used,

$$\hat{\lambda}_{\text{KS}} = \underset{\lambda}{\operatorname{argmin}} \, \sup \, |F_n(\cdot) - \Phi(\cdot)|,$$
$$\hat{\lambda}_{\text{CvM}} = \underset{\lambda}{\operatorname{argmin}} \int_{-\infty}^{\infty} [F_n(\cdot) - \Phi(\cdot)]^2 \, \phi(\cdot),$$

where $F_n(\cdot)$ is the empirical cumulative distribution function estimated by using the normalized residuals, $\Phi(\cdot)$ is the distribution function of a standard normal distribution and $\phi(\cdot)$ its density. The impact of using alternative approaches for estimating λ is studied in a model-based simulation study in Section 2.7.3.

2.5 MSE estimation under transformations

Estimating the MSE of small area estimates is a challenging problem. In the case of the EBP Molina and Rao, 2010 propose a parametric bootstrap procedure following González-Manteiga et al. (2008). In this section we propose two bootstrap schemes for estimating the MSE under transformations. These bootstrap MSE estimators are extended to capture the additional uncertainty due to the estimation of the transformation parameter λ. The difference between the two bootstrap schemes is the mechanism used for generating the bootstrap population. In particular, the first bootstrap generates bootstrap realisations of the random effects and unit-level error terms parametrically. In contrast, the second one is a semi-parametric wild bootstrap which

aims to protect against departures from the assumptions of the model in particular, those of the unit-level error term.
The steps of the proposed parametric bootstrap are as follows:

1. For $b = 1, ..., B$
   (a) Using the sample estimates, $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$, generate $u_i^{(b)} \overset{iid}{\sim} N\left(0, \hat{\sigma}_u^2\right)$ and $e_{ij}^{(b)} \overset{iid}{\sim} N\left(0, \hat{\sigma}_e^2\right)$ and simulate a bootstrap super-population $y_{ij}^{*(b)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(b)} + e_{ij}^{(b)}$.
   (b) Back-transform $y_{ij}^{*(b)}$ to the original scale $y_{ij}^{(b)} = T_\lambda^{-1}\left(y_{ij}^{*(b)}\right)$ and compute the population value of the indicator of interest $I_{i,b}$.
   (c) Extract the bootstrap sample in ${y_{ij}}^{(b)}$ and perform the EBP method, as described in Section 2.4.1. Note, as the back-transformed sample data are used, the transformation parameter λ is re-estimated in each bootstrap replication b.
   (d) Obtain $\hat{I}_{i,b}^{EBP}$.
2. $\widehat{MSE}\left(\hat{I}_i^{EBP}\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2$.

As mentioned before, the proposed parametric bootstrap allows for the additional uncertainty due to the estimation of the transformation parameter. Although the use of an optimal transformation may reduce the deviation from normality, there may still be departures from normality especially in the tails of the distribution of the unit-level error term. To overcome this problem, we propose a semi-parametric bootstrap that relies on the normality of the random effects but generates the unit-level error terms by using the empirical distribution of suitably scaled unit-level residuals. The proposed wild bootstrap scheme is described below:

1. Fit the model 2.1 using an appropriate transformation $T(y_{ij}) = y_{ij}^*$ and obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$.
2. Calculate the sample residuals by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} - \hat{u}_i$.
3. Scale and center the residuals using $\hat{\sigma}_e$. The scaled and centered residuals are denoted by $\hat{\epsilon}_{ij}$.
4. For $b = 1, ..., B$
   (a) Generate $u_i^{(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$.
   (b) Calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(b)}$.
   (c) Match $\eta_{ij}^{(b)}$ with the set of estimated linear predictors $\{\hat{\eta}_k | \eta \in n\}$ 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 $y_{ij}^{*(b)} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + u_i^{(b)} + w_k |\hat{\epsilon}_{\tilde{k}}^{(b)}|$.

(f) Back-transform $T(y_{ij}{}^{*(b)})$ to the original scale and compute the population value $I_{i,b}$.

(g) Extract the bootstrap sample in $y_{ij}{}^{(b)}$ and use the EBP method, as described in Section 2.4.

(h) Obtain $\hat{I}_{i,b}^{EBP}$.

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

The performance of both MSE estimators is compared in a model-based simulation study in Section 2.7.

2.6 The Guerrero case study: Application of data-driven transformations

The benefits of using the proposed EBP approach with data-driven transformations for estimating deprivation and inequality indicators are illustrated in an application using the household data from the ENIGH survey 2010 and the National Population and Housing Census 2010 we introduced in Section 2.3. The aim is to estimate the head count ratio (HCR) and the poverty gap (PGAP) as well as the income quintile share ratio (QSR) for the 81 municipalities in Guerrero. As the ENIGH survey and the census contains only contain information on household level we estimate the poverty and inequality indicators for households and not individuals.

The indicators HCR and PGAP are special cases of the Foster-Greer-Thorbecke (FGT) indicators (Foster et al., 1984) and they depend on a poverty line t which is equal to 0.6 times the median of the target variable. The FGT index of type α for an area i is defined by

$$F_i(\alpha, t) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{t - y_{ij}}{t}\right)^{\alpha} \mathbb{I}(y_{ij} \leq t), \quad \text{for } \alpha = 0, 1, 2,$$

where $\mathbb{I}(\cdot)$ denotes an indicator function which returns 1 if $(\cdot)$ holds and 0 otherwise. When $\alpha = 0$, $F_i(\alpha, t)$ is the HCR and represents the proportion of the households whose income is below the poverty line t. Taking $\alpha = 1$, $F_i(\alpha, t)$ defines the PGAP which is a measure of poverty intensity and quantifies the degree, to which the average income of people living under the poverty line differs from the poverty line. In addition to the two deprivation indicators, we

investigate inequality by the QSR defined by

$$\text{QSR}_i = \frac{\sum_{j=1}^{N_i} \mathbb{I}\left(y_{ij} \geq \mathbf{y}_{0.8}\right) y_{ij}}{\sum_{j=1}^{N_i} \mathbb{I}\left(y_{ij} \leq \mathbf{y}_{0.2}\right) y_{ij}},$$

where $\mathbf{y}_{0.8}$ and $\mathbf{y}_{0.2}$, denote the 80% and 20% quantiles of the target variable respectively. The QSR is a widely used inequality indicator due to its simplicity and straightforward interpretation (Eurostat, 2004). The estimation of the QSR is challenging characterized by a large variability also in large samples. However, we have decided to report the estimated QSR in this paper only for illustrative purposes. In particular, we are interested in showing the increasing importance of the model assumptions when the target parameter depends of the tails of the data distribution and the importance of using data transformation parameters.

Before focusing on the state of Guerrero, we briefly illustrate the need for data-driven transformations in different states in Mexico. Figure 2.2 represents the estimated data-driven Box-Cox transformation parameters $\hat{\lambda}$ (by REML) for each state in Mexico. These estimates vary between 0.13 and 0.37, showing the adaptive feature of data-driven transformations for each state in Mexico. Furthermore, we observe that a fixed logarithmic transformation is not suitable for any of the states.

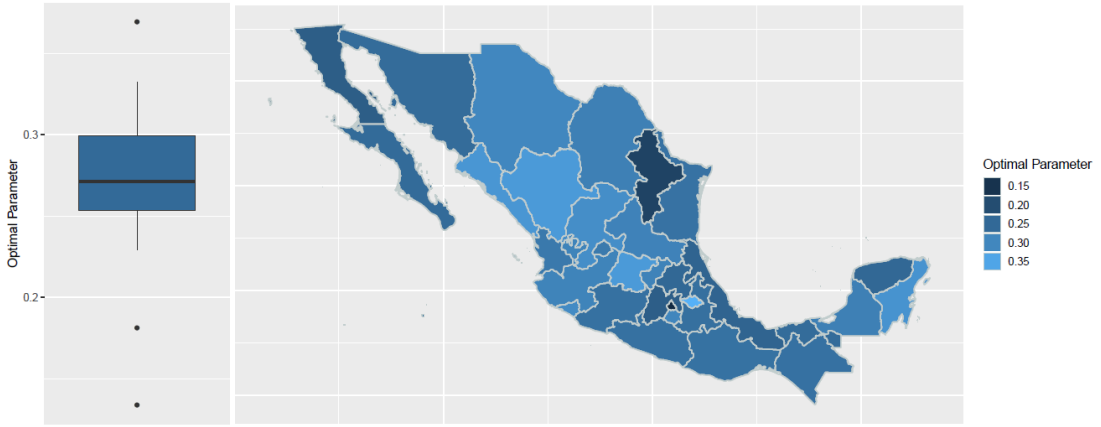


Figure 2.2: Estimated transformation parameters of the Box-Cox transformation in the different states of Mexico

2.6.1 Model checking and residual diagnostics

In Section 2.3 we show that the model assumptions of the working model in the state of Guerrero are not met. We now discuss the use of the proposed data-driven transformations for adapting the working model. In particular, we focus on the three data-driven transformations presented in Section 2.4.2, denoted by *Log-Shift*, *Box-Cox* and *Dual* power transformations and their comparison to (a) a model that use a logarithmic transformation (*Log*) and (b) a model that uses the untransformed income variable (*No*).

To start with, Figure 2.3 provides a graphical representation of the REML maximization for

the transformation parameter λ for log-shift, Box-Cox and dual power transformations in the state of Guerrero. In this case the optimal λs are approximately equal to 156.44, 0.20 and 0.23, respectively (cf. Table 2.3).

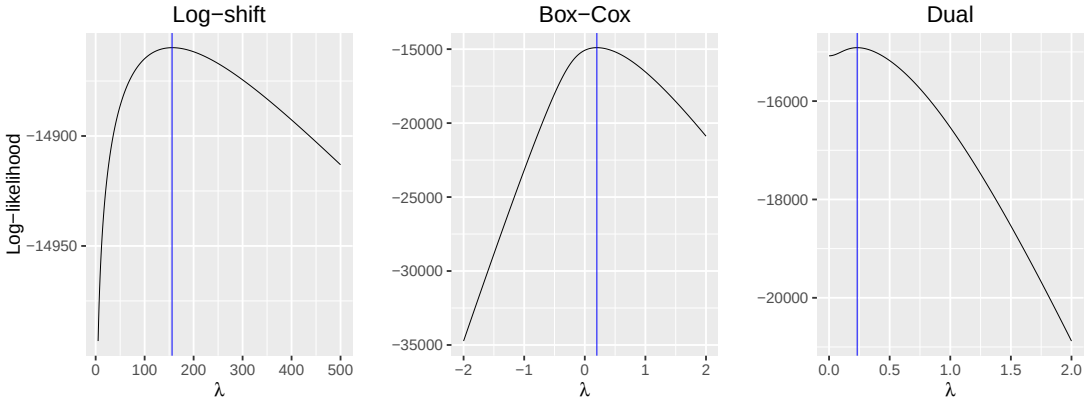


Figure 2.3: Optimal transformation parameter λs for the working model under the log-shift, Box-Cox and dual power transformations in the state of Guerrero

In order to analyze whether the use of transformations improves the predictive power of the model, Table 2.3 reports the percentage of variability explained for each model and its corresponding ICC. As the ICC is larger than 0 in all cases, there appears to be unexplained between area variability and hence the use of the mixed model may be appropriate. Using the untransformed $ictpc$ outcome leads to a marginal (R^2_m) and conditional (R^2_c) coefficients of determination of 0.35 and 0.37, respectively. The use of a logarithmic transformation improves the predictive power of the model in terms of the conditional R^2_c and the marginal R^2_m. However, it can clearly be noted that the use of data-driven transformations increases the predictive power of the model.

Table 2.3: R^2_m, R^2_c, λ_s, and ICC for the working model under the different transformations

	R^2_m	R^2_c	λ	ICC
No	0.351	0.368	-	0.027
Log	0.361	0.458	-	0.151
Log-Shift	0.460	0.522	156.443	0.114
Box-Cox	0.454	0.513	0.199	0.108
Dual	0.454	0.512	0.232	0.106

A detailed analysis of the Gaussian assumptions of the working models corresponding to each transformation is now carried out. The results summarizing the skewness, kurtosis and S-W normality tests are presented in Table 2.4 and the Q-Q plots are presented in Figure 2.4. It should be noted, that at municipal-level, all three data-driven transformations perform similarly and yield good approximations to the normal distribution. In contrast, the household-level residuals show clear departures from normality, especially under the model with a fixed logarithmic transformation and without a transformation. The picture considerably improves for

the data-driven transformations. The log-shift, Box-Cox and dual power transformations lead to very similar results in terms of skewness and kurtosis. We note that the log-shift transformation performs slightly better in terms of kurtosis and skewness compared to the Box-Cox and dual power transformation. These findings are supported by the Q-Q plots displayed in Figure 2.4. The data-driven transformations lead to similar Q-Q plots with more symmetrical and less extreme tails compared to the fixed log transformation. Overall, it appears that the proposed data-driven transformations improve the predictive power of the model and clearly give better approximations to the underlying model assumptions of the linear mixed regression model compared to the use of a fixed logarithmic transformation.

Table 2.4: Skewness, kurtosis and values of the S-W p-values for the municipal- and household-level error terms of the working models for EBP under the different transformations

	Household-level residuals			Municipal-level residuals		
Transformation	Skewness	Kurtosis	p-value	Skewness	Kurtosis	p-value
No	6.338	75.483	0.000	0.448	3.250	0.197
Log	-2.046	16.986	0.000	-1.491	7.059	0.001
Log-Shift	-0.024	4.143	0.000	-0.276	3.485	0.893
Box-Cox	-0.055	6.085	0.000	-0.389	3.861	0.662
Dual	-0.045	6.542	0.000	-0.387	3.889	0.657

2.6.2 Deprivation and inequality indicators for municipalities in Guerrero

Based on the analysis in Section 2.6.1, estimates for the deprivation and inequality indicators presented in Section 2.2 are calculated by using the EBP method under the three data-driven transformations and the fixed logarithmic transformation. MSE estimation is implemented with the wild bootstrap we introduced in Section 2.5 with $B = 500$ bootstrap replications.

Table 2.5 shows summaries over municipalities of point estimates and root MSEs (RMSEs) under the different transformations. In addition we provide a detailed comparison between EBP methods under the different transformations and the direct estimator with corresponding coefficients of variation (CVs) as part of the supplementary material. We observe that the estimates based on the EBP with data-driven transformations are more efficient on average (in terms of RMSE) than the corresponding estimates based on a fixed logarithmic transformation. The effect is especially pronounced for indicators that rely on the tail of the distribution like the QSR. Furthermore, the use of data-driven transformations also has an effect on the point estimates of the indicators. For the HCR and PGAP, the results obtained under the three data-driven transformations are similar to each other, main differences are noticeable when just the fixed logarithmic transformation and no transformation are applied. For instance, the EBP estimates under the model with a logarithmic transformation are on average 5% higher compared to the EBP estimates with data-driven transformations for HCR (see Table 1 in the supplementary material). In addition, the distribution (over municipalities) of the point estimates obtained under EBP with data-driven transformations appear to be closer to the distribution of the direct

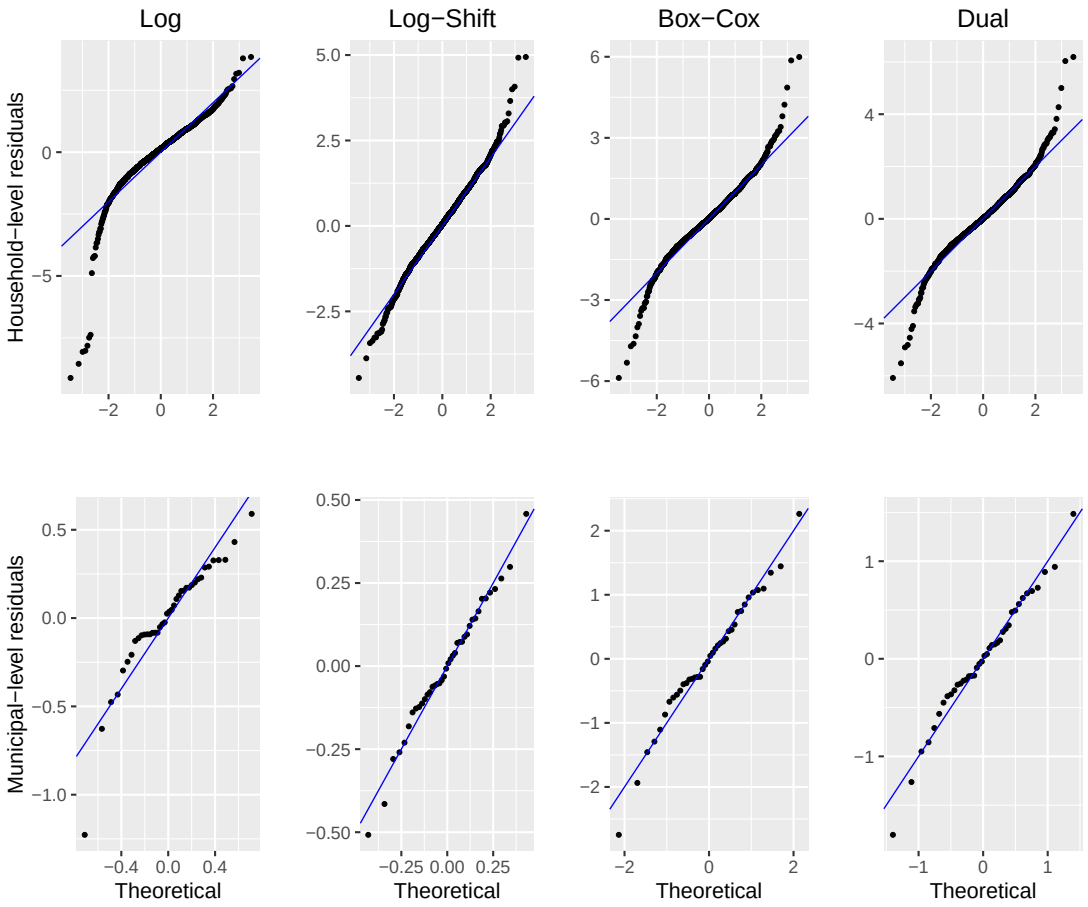


Figure 2.4: Q-Q-plots for the Pearson household-level (upper panels) and municipal-level (lower panels) residuals of the working model for EBP under the different transformations

estimates than the distribution of the EBP under a fixed logarithmic transformation.

Having assessed the estimates from a statistical perspective, we investigate the results in the context of the spatial distribution of poverty and inequality in the state of Guerrero. Figure 2.5 presents the point estimates of HCR, PGAP and QSR at municipal-level. As the point estimates based on the three data-driven transformations are almost identical, we only show the results for the EBP with the log-shift transformation. We observe clear regional differences between the municipalities. Having a closer look to the coastal area in the south-west of Guerrero, where the largest city Acapulco is located, we observe lower levels of poverty (HCR and PGAP) and inequality (QSR) compared to other parts of the state. The coastline to the Pacific Ocean is wealthier due to several tourist destinations like Acapulco, Ixtapa and Zihuatanejo. In contrast, there is also a clear deprivation hotspot in the eastern part of the state Guerrero (e.g. municipalities: Cochoapa el Grande, Metlatnoc and Atlamajalcingo del Monte) with high poverty and inequality rates. These municipalities are home to indigenous populations living in isolated mountain areas.

Table 2.5: Summaries of point estimates and corresponding RMSEs over municipalities in Guerrero

Point Estimation	HCR		PGAP		QSR	
Transformation	Mean	Median	Mean	Median	Mean	Median
Log	0.48	0.49	0.24	0.23	18.03	17.82
Log-Shift	0.44	0.44	0.21	0.20	15.56	14.39
Box-Cox	0.44	0.44	0.22	0.22	16.78	16.39
Dual	0.44	0.44	0.22	0.22	17.38	16.98
RMSE	HCR		PGAP		QSR	
	Mean	Median	Mean	Median	Mean	Median
Log	0.11	0.13	0.07	0.08	32.19	27.54
Log-Shift	0.09	0.09	0.06	0.05	4.68	1.91
Box-Cox	0.09	0.10	0.06	0.06	2.98	2.95
Dual	0.09	0.10	0.06	0.06	2.85	2.76

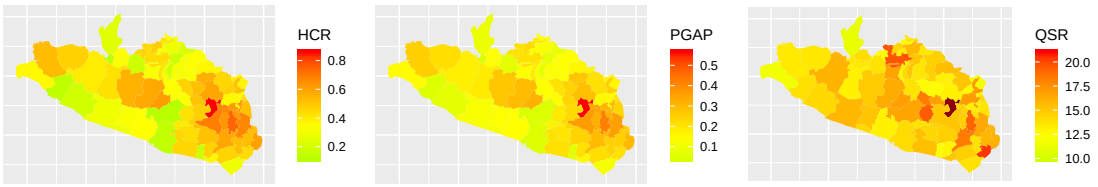


Figure 2.5: Maps of the HCR, PGAP and QSR in Guerrero for the EBP method under the log-shift transformation at municipal-level

2.7 Model-based simulation study

In this section, we present results from a model-based simulation study that aims to evaluate the performance of the proposed methods. In Section 2.7.1 we analyze the behaviour of the data-driven transformation parameter under four scenarios for the distributions of the area and unit-level error terms. In Section 2.7.2 we investigate the ability of the proposed methods to provide more precise small area estimates than the EBP with a fixed logarithmic transformation or without a transformation and assess the performance of the proposed MSE estimators. Finally, in Section 2.7.3 we evaluate the methods for estimating the transformation parameter. In addition we have also conducted a design-based simulation study with a variable available in the census data that is highly correlated with the target variable ($ictpc$) in the application. The results are provided as part of the supplementary material.

We generate finite populations U of size $N = 10000$, partitioned into $D = 50$ areas $U_1, U_2, \ldots,$ U_D of sizes $N_i = 200$. The samples are selected by a stratified random sampling with strata defined by the 50 small areas. This leads to a sample size of $n = \sum_{i=1}^{D} n_i = 921$ whereby the area-specific sample sizes n_i vary between 8 and 29. We chose the sample sizes mainly because of two reasons. First, we want to assess the data-driven transformations under extreme but realistic cases. Second, the sample sizes are similar in the case study.

Four scenarios, denoted by *Normal*, *Log-scale*, *Pareto* and *GB2*, are considered. Details about the data generating mechanisms of the different scenarios are provided in Table 2.6. Under scenario *Normal*, data are generated by using Normal distributions for the random effects and unit-level errors. Under the second scenario random effects and unit-level errors are generated under a log-normal distribution such that a fixed logarithmic transformation is suitable. Scenarios *Pareto* and *GB2* are settings that attempt to replicate realistic situations for income data. In particular, these distributions mimic the features of income-based variables, namely unimodal, leptokurtic, and highly skewed data influenced by outliers. Random effects are generated by using a Normal distribution and unit-level error terms are generated under a *Pareto* and *GB2* scenario respectively. Each setting was repeated independently $M = 500$ times. We focus on the three data-driven transformations, namely log-shift, Box-Cox, and dual power transformations, and compare these to the case of a fixed logarithmic transformation and the case of using untransformed data.

Table 2.6: Model-based simulation settings for the analysis of the MSE

Scenario	Model	x_{ij}	z_{ij}	μ_i	u_i	e_{ij}
Normal	$4500 - 400x_{ij} + u_i + e_{ij}$	$N(\mu_i, 3)$	-	$U[-3, 3]$	$N\left(0, 500^2\right)$	$N(0, 1000^2)$
Log-scale	$\exp(10 - x_{ij} - 0.5z_{ij} + u_i + e_{ij})$	$N(\mu_i, 2)$	$N(0, 1)$	$U[2, 3]$	$N\left(0, 0.4^2\right)$	$N(0, 0.8^2)$
Pareto	$12000 - 400x_{ij} + u_i + e_{ij} - \bar{e}$	$N(\mu_i, 7.5)$	-	$U[-3, 3]$	$N\left(0, 500^2\right)$	$\sqrt{2}\text{Pareto}(3, 2000^2)$
GB2	$8000 - 400x_{ij} + u_i + e_{ij} - \bar{e}$	$N(\mu_i, 5)$	-	$U[-1, 1]$	$N\left(0, 500^2\right)$	$\text{GB2}(2.5, 1700, 18, 1.46)$

2.7.1 Behaviour of the data-driven transformation parameters

Figure 2.6 shows box plots of the estimated transformation parameters λ for the log-shift, Box-Cox and dual power transformations (over $M = 500$ replications) under the four simulation settings. The data-driven transformation parameters are estimated by REML. Under the *Normal* setting the parameters of the Box-Cox and dual power transformations are close to one indicating that no transformation is needed. In the *Log-scale* scenario, the data was generated in such a way that normality may be achieved by applying the logarithmic transformation. In this case the log-shift transformation parameter is close to zero and the same holds for the parameters of Box-Cox and dual power transformations. For the other two scenarios (*Pareto* and *GB2*), the data-driven parameters are between 0.25 and 0.5, so neither using a logarithmic transformation nor ignoring the need for a transformation is appropriate. Overall, the results indicate that the data-driven transformations behave as expected in the four scenarios and adapt to the shapes of the data distributions.

2.7.2 Performance of the EBP under data-driven transformations

In this section we compare the performance of the proposed methods to the case of (a) fixed logarithmic transformation and (b) no transformation. We then assess the performance of the MSE estimators. Five estimators of small area deprivation and inequality indicators (HCR, PGAP and QSR) are evaluated. The EBP and the corresponding MSE estimators are implemented using $L = 100$ and $B = 500$. The following quality measures averaged over Monte-Carlo

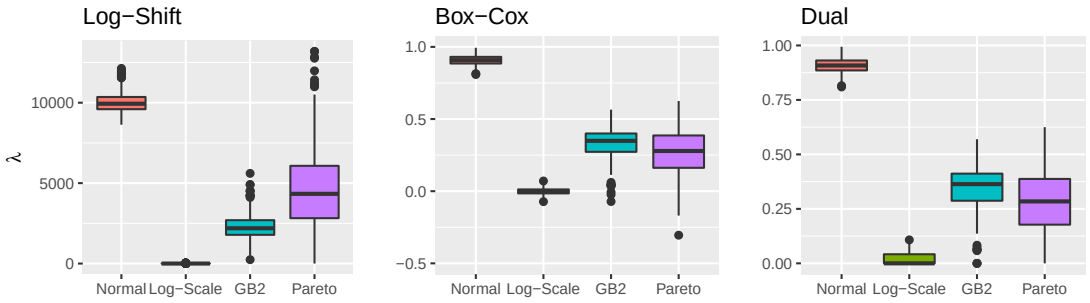


Figure 2.6: Estimated transformation parameters for the log-shift, Box-Cox and dual power transformations under the different settings.

replications M are used to assess the performance of a small area estimator in area i:

$$\text{RMSE}\left(\hat{I}_i^{\text{method}}\right) = \left[\frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}_i^{\text{method}(m)} - I_i^{(m)}\right)^2\right]^{1/2},$$

$$\text{Bias}\left(\hat{I}_i^{\text{method}}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}_i^{\text{method}(m)} - I_i^{(m)}\right),$$

where $\hat{I}_i^{\text{method}}$ denotes the estimated indicator in area i based on any of the five methods under consideration and I_i denotes the corresponding true value in area i. Table 2.7 presents the results split by the four scenarios. It shows median and mean values of RMSE and bias averaged over small areas. Under the *Normal* scenario the EBP without transformation is the gold, but the EBP with data-driven transformations (log-shift, Box-Cox and dual power) perform similarly in terms of RMSE and bias. The same picture emerges in the *Log-scale* scenario where the EBP with a logarithmic transformation is the gold standard, but again the EBP with data-driven transformations perform well both in terms of RMSE and bias. These results confirm our expectations that the EBP with data-driven transformations adapt to the shape of the data distribution. Under the *GB2* and *Pareto* scenarios we notice that the EBP with a fixed transformation or without transformation is inferior to the EBP with data-driven transformations both in terms of RMSE and Bias. The differences are especially pronounced for QSR which is very sensitive to the tails of the distribution. Furthermore, the estimates based on data-driven transformations are almost unbiased or have a small bias. A closer look at the data-driven transformations indicates that EBP with a log-shift transformation performs slightly better than the EBP with Box-Cox and dual power transformations under the *GB2* and *Pareto* scenarios. Overall, it appears that the proposed EBP method with data-driven transformations adapts to the underlying distribution of the data, and hence improves the precision of small area estimates.

Table 2.7: Summaries of estimated RMSEs and Bias over the model-based settings

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Normal							
RMSE	No	0.0338	0.0357	0.0136	0.0154	0.3259	1.2765
	Log-Shift	0.0344	0.0363	0.0155	0.0175	0.3898	0.6710
	Box-Cox	0.0343	0.0358	0.0134	0.0156	0.3348	1.1178
	Dual	0.0343	0.0358	0.0134	0.0156	0.3346	0.5797
BIAS	No	0.0000	0.0007	0.0002	0.0009	0.0049	0.0899
	Log-Shift	0.0029	0.0039	-0.0067	-0.0076	-0.1000	-0.2190
	Box-Cox	0.0016	0.0027	-0.0021	-0.0025	-0.0396	-0.0807
	Dual	0.0016	0.0027	-0.0021	-0.0024	-0.0458	-0.1193
Log-Scale							
RMSE	Log	0.0583	0.0605	0.0358	0.0367	4.9100	4.8969
	Log-Shift	0.0583	0.0605	0.0358	0.0367	4.9024	4.8985
	Box-Cox	0.0581	0.0604	0.0358	0.0367	4.9731	4.9717
	Dual	0.0584	0.0605	0.0359	0.0367	4.9025	4.9093
BIAS	Log	-0.0011	-0.0009	-0.0007	-0.0003	0.0394	0.1143
	Log-Shift	-0.0020	-0.0017	-0.0011	-0.0007	-0.0873	-0.0072
	Box-Cox	-0.0009	-0.0006	-0.0008	-0.0004	0.1499	0.2106
	Dual	-0.0024	-0.0021	-0.0009	-0.0005	-0.1610	-0.0992
GB2							
RMSE	No	0.0650	0.0656	0.0552	0.0552	17.7364	32.0686
	Log	0.0912	0.0908	0.0272	0.0270	1.8979	1.9002
	Log-Shift	0.0418	0.0415	0.0127	0.0132	0.4286	0.4411
	Box-Cox	0.0471	0.0469	0.0136	0.0139	0.4708	0.4753
	Dual	0.0472	0.0470	0.0137	0.0140	0.4715	0.4760
BIAS	No	0.0471	0.0477	0.0481	0.0479	1.8355	2.0825
	Log	0.0746	0.0747	0.0169	0.0169	1.4718	1.4692
	Log-Shift	0.0176	0.0179	-0.0008	-0.0013	0.0546	0.0523
	Box-Cox	0.0274	0.0274	0.0035	0.0031	0.1780	0.1721
	Dual	0.0275	0.0274	0.0037	0.0034	0.1800	0.1747
Pareto							
RMSE	No	0.0448	0.0444	0.0622	0.0613	1.6814	3.6057
	Log	0.0304	0.0306	0.0082	0.0084	0.3887	0.3994
	Log-Shift	0.0185	0.0196	0.0060	0.0063	0.1661	0.1779
	Box-Cox	0.0192	0.0202	0.0059	0.0062	0.1786	0.1901
	Dual	0.0192	0.0203	0.0059	0.0062	0.1782	0.1902
BIAS	No	0.0277	0.0287	0.0166	0.0160	0.3173	0.3132
	Log	0.0086	0.0081	-0.0030	-0.0037	0.2068	0.2034
	Log-Shift	0.0003	-0.0001	-0.0034	-0.0041	0.0305	0.0300
	Box-Cox	0.0030	0.0026	-0.0031	-0.0037	0.0525	0.0530
	Dual	0.0030	0.0027	-0.0031	-0.0037	0.0522	0.0530

We now turn our attention to the performance of the MSE estimators. We denote by *parametric* and *wild* the proposed parametric bootstrap and proposed semi-parametric wild bootstrap respectively. The aim of this part is twofold. Firstly, we assess the performance of the two proposed MSE estimators we introduced in Section 2.5. Secondly, we investigate the ability of the wild bootstrap to protect against departures from the assumptions of the unit-level error term. Starting with the first aim, Table 2.8 reports the results for the two MSE estimators and presents the mean and median values of relative RMSE and relative bias -over Monte-Carlo replications and areas- of the EBP with Box-Cox transformation. For calculating the RMSE and relative bias we treat the empirical MSE (over Monte-Carlo replications) as the true MSE. The results for the EBP with a log-shift transformation and dual power transformation are very similar and although they are omitted they are available on request from the authors.

Table 2.8: Performance of MSE estimators in model-based simulations: EBP with Box-Cox transformation

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Normal							
rel. RMSE[%]	Parametric	8.30	9.22	9.15	9.47	15.25	21.23
	Wild	14.57	14.77	14.21	14.61	17.46	20.93
rel. Bias[%]	Parametric	6.64	7.27	-1.17	-0.12	-7.72	-12.61
	Wild	8.05	8.04	2.17	3.23	-1.01	-1.46
Log-Scale							
rel. RMSE[%]	Parametric	11.14	12.00	19.19	19.57	19.10	19.75
	Wild	16.82	17.00	22.70	22.95	25.34	25.62
rel. Bias[%]	Parametric	6.10	6.29	5.70	6.36	7.91	7.92
	Wild	7.69	7.82	7.34	7.39	6.58	6.78
GB2							
rel. RMSE[%]	Parametric	21.71	21.86	20.89	20.57	43.75	43.58
	Wild	19.01	19.39	14.76	15.12	26.21	27.23
rel. Bias[%]	Parametric	-20.04	-19.74	-16.88	-15.92	-42.90	-42.74
	Wild	-14.59	-14.64	-5.45	-5.75	-21.72	-22.53
Pareto							
rel. RMSE[%]	Parametric	11.31	12.60	35.60	34.78	50.04	51.63
	Wild	26.18	28.44	23.58	26.04	28.60	33.40
rel. Bias[%]	Parametric	2.43	3.38	-33.82	-31.16	-49.51	-51.06
	Wild	19.21	21.37	-8.28	-3.28	-23.02	-26.79

We note that, on average, the proposed *parametric* and *wild* bootstrap approaches for the EBP with a Box-Cox transformation have small positive relative bias (HCR and PGAP indicators) in the *Normal* and *Log-scale* settings. However, *parametric* bootstrap shows some underestimation in the case of QSR. In this latter case *wild* bootstrap appears to be associated with smaller relative bias. For the *Normal* and *Log-scale* scenarios *parametric* bootstrap also has

smaller relative RMSE than *wild* bootstrap. Nevertheless, *wild* bootstrap provides reasonable results for HCR and PGAP and reduces the underestimation for QSR. When the distributional assumptions are not met, as in the (*GB2* and *Pareto*) scenarios, *parametric* tends to clearly underestimate the MSE (except for the HCR in the *Pareto* scenario). Although *wild* bootstrap does not completely eliminate this bias, it greatly reduces it and provides more stable MSE estimates in terms of relative RMSE. These results indicate that departures from the model assumptions -even after using data transformations- can impact MSE estimation with parametric methods. The problem is more pronounced when estimating parameters that depend on the tails of the distribution as is the case with the QSR. In those cases, the use of semi-parametric bootstrap, at least as a supplementary MSE estimation method, can offer some protection against misspecification.

2.7.3 Impact of alternative estimation methods for λ

In this last section we explore the use of non-parametric alternatives to the REML approach for estimating data-driven transformation parameters (see Section 2.4.3). Here, we study five estimation methods. These are the REML approach, the minimization of the skewness (*Skew*) and the pooled skewness (*poolSkew*), and the distance-based criteria Kolomogorov-Smirnov (*KS*) and Cramér-von Mises (*CvM*) we introduced in Section 2.4.3.

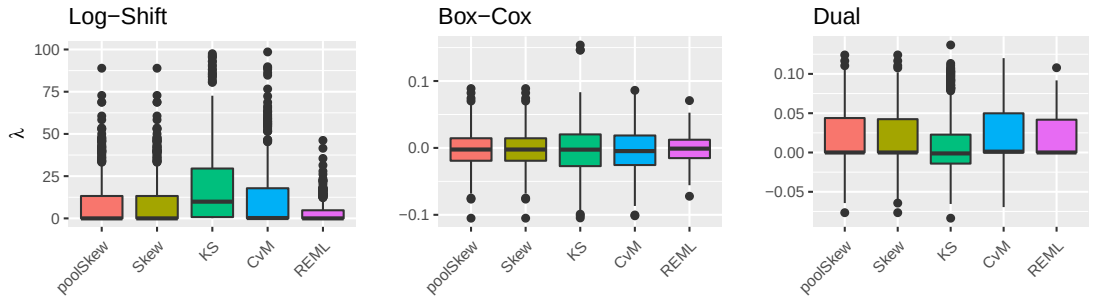


Figure 2.7: Box-plots of estimated transformation parameters for the log-scale scenario using different estimation methods

The five methods estimate transformation parameters close to the theoretically correct ones, in the scenarios those are known. For instance, in the Log-scale scenario, the estimated transformation parameters under the different estimation methods are shown in Figure 2.7 and Table 2.9. We observe that although the five methods provide similar estimates of λ, the REML method has smaller variability. In our model-based simulations we further studied the impact of the estimation method of the transformation parameter on point and MSE estimation and we conclude that this only marginally influences the quality of small area estimates. These results are available from the authors upon request.

Overall, these results suggest that for the scenarios we considered in this paper the method used to estimate the transformation parameter does not have a noticeable impact on small area estimation and REML appears to be the most stable method.

Table 2.9: Mean and median of estimated transformation parameters under the log-scale scenario using different estimation methods

	Log-Shift		Box-Cox		Dual	
	Mean	Median	Mean	Median	Mean	Median
poolSkew	9.381	0.000	-0.002	-0.002	0.016	0.000
Skew	9.381	0.000	-0.002	-0.002	0.015	0.000
KS	23.906	10.816	-0.003	-0.003	0.009	-0.001
CvM	11.954	0.211	-0.004	-0.005	0.025	0.001
REML	3.349	0.000	-0.002	-0.001	0.021	0.000

2.8 Conclusions and future research directions

In this paper we investigate data-driven transformations for small area estimation. In particular, we propose an EBP approach with data-driven transformations estimated with likelihood-based methods. The use of scaled transformations (conditional on the Jacobian) allows for the use of standard software for fitting the linear mixed regression model. Three types of transformations are discussed log-shift, Box-Cox and dual power transformations. We further explore the use of parametric and semi-parametric wild bootstrap for MSE estimation that also captures the uncertainty from estimating the data driven transformation parameter. Semi-parametric bootstrap is used for protecting against departures from the model assumptions. Model-based simulations demonstrate the ability of the proposed EBP method to adapt to the shape of the data distribution and hence provide more efficient estimates than a fixed logarithmic transformation or the case where no transformation is used. Although the paper focuses on the EBP the proposed methods are applicable to other small area estimators for example, the ELL approach (Elbers, Lanjouw, et al., 2003). The methods proposed in this paper can be implemented by using the R Package **emdi** (Kreutzmann, Marek, et al., 2019). The package supports the user by estimating and mapping regionally disaggregated indicators. Although this package already includes the logarithm and Box-Cox transformations, some research effort should be shifted towards the development of relevant software which includes in more detail the use of data-driven transformations in the SAE context.

Further research can investigate the use of multiparameter transformation families. This may allow for better control of higher moments and hence better adaptation to the distribution of the data. Since likelihood-based approaches might be influenced by outliers, it would be also interesting to investigate robust estimation methods. Model selection with data driven transformations presents additional challenges. Finding a good working model depends on the method of transformation. In this paper we first find a working model and keep this fixed when considering different data-driven transformations. However, this may not offer the best approach to model selection. Approaches that simultaneously consider both steps for linear regression models have been proposed (Laud and Ibrahim, 1995; Hoeting and Ibrahim, 1998; Hoeting, Raftery, et al., 2002). Extending these approaches to the case of linear mixed models is an

open research problem. Finally, comparing the EBP with data-driven transformations to EBP approaches with alternative parametric assumptions (Diallo and Rao, 2014; Graf et al., 2019) is empirical work that remains open.

Acknowledgements

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 authors are grateful to CONEVAL for providing the data used in empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are only produced for illustrating the methods. The authors are grateful for the computation time provided by the HPC service of the Freie Universität Berlin.

2.9 Appendix

2.9.1 Derivation of scaled transformations

In this appendix we derive the Jacobian and the corresponding scaling factors presented in Table 2.2 for the log-shift, Box-Cox, and dual power transformations.

Log-shift transformation

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the log-shift transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D} \prod_{j=1}^{n_i} \left| \frac{dy_{ij}^*(\lambda)}{dy_{ij}} \right| \\ &= \prod_{i=1}^{D} \prod_{j=1}^{n_i} (y_{ij} + \lambda)^{-1}. \end{aligned}$$

The log-likelihood function in (2.3) can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}]^T \boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}] - n\log \underbrace{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+\lambda)\right)^{\frac{1}{n}}}_{=\bar{y}_\lambda}. \end{aligned}$$

In order to obtain the scaled log-shift transformation, $z_{ij}^*(\lambda)$, the denominator of the term

$\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given by:

$$\begin{aligned} 1/J(\lambda,\boldsymbol{y})^{\frac{1}{n}} = J(\lambda,\boldsymbol{y})^{-\frac{1}{n}} &= \left[\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+\lambda)^{-1}\right]^{-\frac{1}{n}} \\ &= \bar{y}_\lambda. \end{aligned}$$

Therefore, the scaled log-shift transformation is defined as follows:

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}} = \bar{y}_\lambda \, \log(y_{ij}+\lambda)$$

for $y_{ij} > -\lambda$.

Box-Cox transformation

Let $J(\lambda,\boldsymbol{y})$ be the Jacobian of the Box-Cox transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda,\boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| \\ &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)^{\lambda-1}. \end{aligned}$$

The log-likelihood function in (2.3) can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y},\lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda)-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda)-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}] \\ &+ n(\lambda-1)\log\underbrace{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)\right)^{\frac{1}{n}}}_{=\bar{y}}. \end{aligned}$$

In order to obtain the scaled transformation of the Box-Cox family, $z_{ij}^*(\lambda)$, the denominator of the term $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given by:

$$\begin{aligned} 1/J(\lambda,\boldsymbol{y})^{\frac{1}{n}} = J(\lambda,\boldsymbol{y})^{-\frac{1}{n}} &= \left[\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)^{\lambda-1}\right]^{-\frac{1}{n}} \\ &= \bar{y}^{-(\lambda-1)}. \end{aligned}$$

Therefore, the scaled Box-Cox transformation is defined as follows:

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda, \boldsymbol{y})^{1/n}} = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\bar{y}^{\lambda-1}\lambda}, & \lambda \neq 0, \\ \bar{y}\log(y_{ij}+s), & \lambda = 0, \end{cases}$$

for $y_{ij} > -s$.

Dual power transformation

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the dual power transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| \\ &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1} + (y_{ij}+s)^{-\lambda-1}}{2}. \end{aligned}$$

The log-likelihood function in (2.3) can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}] \\ &+ n\log\underbrace{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1} + (y_{ij}+s)^{-\lambda-1}}{2}\right)^{\frac{1}{n}}}_{=\bar{y}_\lambda}. \end{aligned}$$

In order to obtain the scaled dual transformation, $z_{ij}^*(\lambda)$, the denominator of the term $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given by:

$$\begin{aligned} 1/J(\lambda, \boldsymbol{y})^{1/n} = J(\lambda, \boldsymbol{y})^{-\frac{1}{n}} &= \left[\prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1} + (y_{ij}+s)^{-\lambda-1}}{2}\right]^{-\frac{1}{n}} \\ &= \bar{y}_\lambda^{-1}. \end{aligned}$$

Therefore, the scaled dual transformation is defined as follows:

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda, \boldsymbol{y})^{1/n}} = \begin{cases} \bar{y}_\lambda^{-1}\frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \bar{y}_\lambda^{-1}\log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

for $y_{ij} > -s$.

2.10 Supplementary material

2.10.1 The Guerrero case study: Additional results

As part of the supplementary material we present a comparison between EBP methods under the different transformations and the direct estimator. Tables 2.10, 2.11 and 2.12 show summaries over in-sample municipalities in Guerrero of point estimates, estimated RMSEs and CVs for HCR, PGAP and QSR, respectively. The results show that the EBP with data-driven transformations are more efficient (in terms of RMSEs and CVs) than the corresponding direct estimates and the estimates produced with the logarithmic model and the untransformed model. When comparing direct and model-based estimates we observe that the direct estimates are less shrunk compared to model-based estimates with data-driven and logarithmic transformations. In addition, the distribution (over municipalities) of the point estimates obtained under EBP with data-driven transformations appear to be closer to the distribution of the direct estimates than the distribution of the EBP under a fixed logarithmic transformation.

Point Estimation			HCR			
Transformation	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.06	0.20	0.42	0.42	0.61	0.90
No	0.14	0.30	0.39	0.38	0.47	0.54
Log	0.16	0.33	0.49	0.46	0.55	0.85
Log-Shift	0.12	0.29	0.44	0.41	0.52	0.74
Box-Cox	0.12	0.30	0.45	0.41	0.53	0.75
Dual	0.12	0.30	0.45	0.41	0.52	0.74
RMSE						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.02	0.06	0.08	0.08	0.10	0.12
No	0.04	0.06	0.08	0.09	0.11	0.15
Log	0.04	0.05	0.06	0.07	0.08	0.13
Log-Shift	0.01	0.05	0.06	0.06	0.07	0.09
Box-Cox	0.02	0.05	0.06	0.06	0.07	0.09
Dual	0.01	0.05	0.06	0.06	0.07	0.09
CV						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.07	0.13	0.22	0.28	0.34	0.98
No	0.14	0.19	0.22	0.23	0.26	0.54
Log	0.08	0.11	0.14	0.17	0.19	0.35
Log-Shift	0.07	0.12	0.13	0.16	0.17	0.41
Box-Cox	0.07	0.11	0.13	0.16	0.17	0.41
Dual	0.07	0.11	0.13	0.16	0.17	0.41

Table 2.10: Summaries of point estimates, estimated RMSEs and CVs for HCR over in-sample municipalities in Guerrero

Point Estimation	PGAP					
Transformation	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.00	0.08	0.21	0.21	0.32	0.78
No	0.26	0.67	0.90	0.91	1.20	1.46
Log	0.06	0.14	0.23	0.23	0.29	0.72
Log-Shift	0.04	0.12	0.21	0.20	0.26	0.56
Box-Cox	0.05	0.13	0.22	0.22	0.28	0.60
Dual	0.05	0.14	0.22	0.22	0.28	0.60
RMSE						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.00	0.03	0.05	0.05	0.06	0.08
No	0.13	0.38	0.48	0.46	0.59	0.65
Log	0.02	0.03	0.04	0.04	0.05	0.12
Log-Shift	0.01	0.02	0.04	0.03	0.04	0.06
Box-Cox	0.01	0.03	0.04	0.04	0.05	0.07
Dual	0.01	0.03	0.04	0.04	0.05	0.08
CV						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.05	0.17	0.29	0.33	0.38	0.99
No	0.28	0.47	0.51	0.52	0.53	0.90
Log	0.10	0.16	0.19	0.23	0.24	0.57
Log-Shift	0.09	0.13	0.16	0.21	0.22	0.66
Box-Cox	0.09	0.15	0.17	0.21	0.22	0.67
Dual	0.09	0.14	0.17	0.21	0.21	0.66

Table 2.11: Summaries of point estimates, estimated RMSEs and CVs for PGAP over in-sample municipalities in Guerrero

Point Estimation			QSR			
Transformation	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	5.11	7.22	9.37	12.22	14.24	46.86
No	-98.30	-4.76	-2.74	-6.33	-2.00	-1.44
Log	13.63	16.74	19.10	19.22	21.90	26.51
Log-Shift	11.26	13.73	14.82	15.02	16.15	19.64
Box-Cox	12.12	15.15	16.61	16.75	18.69	23.05
Dual	12.28	15.45	17.07	17.25	19.46	23.93
RMSE						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.90	1.92	3.23	9.50	4.59	173.74
No	2.55	31.26	206.15	828.75	776.01	6538.94
Log	8.21	18.60	24.70	27.84	37.51	65.44
Log-Shift	1.17	1.51	1.92	3.67	4.25	14.28
Box-Cox	1.16	2.20	2.70	2.69	3.31	4.64
Dual	0.92	1.86	2.40	2.44	3.04	4.50
CV						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Direct	0.11	0.24	0.31	0.56	0.45	5.97
No	-2252.40	-207.72	-47.970	-193.43	-11.47	-1.57
Log	0.37	0.85	1.24	1.59	2.03	4.71
Log-Shift	0.08	0.10	0.13	0.25	0.33	1.04
Box-Cox	0.09	0.13	0.16	0.16	0.19	0.25
Dual	0.06	0.11	0.14	0.14	0.17	0.23

Table 2.12: Summaries of point estimates, estimated RMSEs and CVs for QSR over in-sample municipalities in Guerrero

2.10.2 Design-based simulation study

As part of the supplementary materials a design-based simulation study for Guerrero is carried out. The aim of such a study, is to investigate and compare the behaviour of the estimation methods presented in this paper under realistic conditions. As mentioned in the paper, the total household per capita income *(ictpc)*, which is available only in the survey data, is used as the outcome in the application. The earned income from work per capita (*inglabpc*), is a variable that is highly correlated with the target variable and is available for the population of Guerrero in the census. The design-based simulation is therefore based on repeated sampling from the National Population and Housing Census 2010 and the census variable *inglabpc* is used as the proxy. The results presented in Table 2.13 split by the in-sample and out-of-sample municipalities. The table reports median and mean values of the RMSE and bias of the predictors (*Direct, No, Log-shift, Box-Cox* and *Dual*) averaged over municipalities. These estimators of small area deprivation and inequality indicators (HCR, PGAP and QSR) are evaluated. From the fixed pseudopopulation, $T = 500$ samples are independently drawn following a sampling design similar to the Household Income and Expenditure Survey (ENIGH) 2010. As in the ENIGH survey, the sample size is 1611 households with 40 in-sample- and 41 out-of-sample municipalities.

According to the results from the design-based simulation study, the use of data-driven transformations under the EBP method provides better estimates in terms of bias and efficiency for all indicators than in case no transformation or the logarithm is used. This analysis confirms the results obtained in the model-based simulation study presented in this paper.

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
In-sample							
RMSE	Direct	0.10	0.10	0.06	0.06	8.98	14.05
	No	0.07	0.08	0.77	0.79	18.88	23.89
	Log	0.07	0.07	0.04	0.04	3.79	8.30
	Log-Shift	0.05	0.05	0.03	0.04	2.19	6.50
	Box-Cox	0.05	0.05	0.04	0.04	2.29	6.46
	Dual	0.05	0.05	0.04	0.04	2.33	6.47
Bias	Direct	0.00	0.00	0.00	0.00	-0.85	-0.38
	No	-0.03	-0.03	0.74	0.76	-18.74	-23.46
	Log	0.05	0.05	0.02	0.02	0.78	-3.43
	Log-Shift	0.01	0.01	0.00	-0.00	-0.95	-4.60
	Box-Cox	0.01	0.01	0.01	0.01	0.29	-3.98
	Dual	0.01	0.01	0.01	0.01	0.46	-3.82
Out-of-sample							
RMSE	No	0.06	0.08	0.83	0.87	19.54	22.73
	Log	0.05	0.07	0.04	0.05	4.81	8.08
	Log-Shift	0.05	0.06	0.04	0.05	4.35	8.16
	Box-Cox	0.05	0.06	0.04	0.05	4.26	7.42
	Dual	0.05	0.06	0.04	0.05	4.46	7.43
Bias	No	-0.05	-0.06	0.81	0.84	-19.54	-22.73
	Log	0.05	0.04	0.02	0.02	-0.12	-4.03
	Log-Shift	-0.00	-0.01	-0.00	-0.01	-0.33	-4.38
	Box-Cox	0.01	0.00	0.01	0.01	0.50	-3.59
	Dual	0.01	0.00	0.01	0.01	0.67	-3.40

Table 2.13: Performance of predictors over municipalities in Guerrero in design-based simulations

Chapter 3

A Framework for Producing Small Area Estimates Based on Area-Level Models in R

3.1 Introduction

Small area estimation (SAE) has gained importance not only in research but also in many fields of application to get a better insight of indicators at a small-scale level. Among others, SAE is used for estimating socio-economic measures like income, poverty and health or indicators for agriculture (Datta, Fay, et al., 1991; Tzavidis, Chambers, et al., 2012; Zhang et al., 2015; Pratesi, 2016). Especially official statistics and economic or political decision makers benefit from reliable estimation of disaggregated indicators and thus SAE methods. Existing surveys were often not planned for these disaggregated levels and show only small sample sizes which often leads to a low precision of the estimates. SAE methods can be employed to avoid expensive and time-consuming enlargements of the sample size of surveys. The idea is to combine data sources with model-based approaches. Existing survey data will be enriched by auxiliary information, e.g., from census or register data, to improve the accuracy of the estimation of the indicators on area- or domain- level. The terms area and domain can be used interchangably and refer either to a geographic area or to any subpopulation of a population of interest like socio-demographic groups. Among others, Pfeffermann (2013), Rao and Molina (2015), Tzavidis, Zhang, et al. (2018) and Jiang and Rao (2020) give comprehensive overviews of SAE methods.

The main goal of the package **emdi** is the simplification of estimating these regionally disaggregated indicators. The package version 1.1.7 contains direct estimation based exclusively on survey data and model-based estimation using the unit-level empirical best predictor (EBP) method (Molina and Rao, 2010). The EBP approach is powerful since it enables the simul-

taneous estimation of various indicators. For this, it relies on unit-level information, i.e., information about each unit in each domain. Even though survey data often provides unit-level information, access to census or register data at unit-level is less likely. Hence, area-level models provide a valuable alternative. First, only area-level aggregates are needed for the estimation of the regional indicators. Second, area-level models can consider the survey design by integrating the sampling weights. Third, the computation is faster compared to the computational intensive EBP approach.

Various R packages that employ different area-level models are available on the Comprehensive R Archive Network (CRAN): The package **smallarea** (Nandy, 2015) offers different variance estimation methods (maximum likelihood (ML), residual maximum likelihood (REML), Prasad-Rao- and Fay-Herriot method-of-moment) for the standard Fay-Herriot (FH) model and a function to estimate unknown sampling variances. The opportunity of estimating unit- and area-level models under heteroscedasticity is provided by the **JoSAE** package (Breidenbach, 2015). The package **saery** (Lefler et al., 2014) provides functions for the estimation of temporal FH models. The robust estimation of area-level models with spatial and/or temporal structures in the random effects is supported by package **saeRobust** (Warnholz, 2018). The estimation of multivariate FH models is possible with package **msae** (Permatasari and Ubaidillah, 2020). The package **hbsae** (Boonstra, 2012) allows for the fitting of unit- and area-level models by frequentist or hierarchical Bayesian approaches. The possibility of estimating FH models and some of its extensions in a Bayesian framework is also given by the **BayesSAE** package (Shi, 2018). Further on, the **mme** package (Lopez-Vizcaino et al., 2019) allows the building of Gaussian area-level multinominal mixed-effects models in the SAE context. One of the commonly used packages is the **sae** package (Molina and Marhuenda, 2015). It includes a wide range of area-level models (the standard FH model with REML, ML and FH method-of-moment model fitting and a spatial and a spatio-temporal extension of the FH model) and unit-level models (the nested error linear regression model of Battese et al. (1988) and the EBP approach). Table 3.1 gives an overview of the packages and the implemented methodology. Package **emdi** version 2.0.1 expands the existing packages for the following reasons:

- None of the existing packages contains such a variety of different area-level models.
- In addition to the spatial and robust area-level models that are already available in existing R packages, **emdi** includes also area-level models that are not available in existing packages: adjusted variance estimation methods and transformation options for the standard FH model, and a measurement error FH model.
- Package **emdi** offers user-friendly tools that go beyond model estimation for the new and existing methods like specific diagnostic tools both in form of a summary and graphical diagnostics, and the comparison of the model-based with direct estimates and their respective mean squared error (MSE) estimates. Furthermore, benchmarking options, geographically visualization of the results in form of high quality maps, and export of

	Package									
Area-level model	***smallarea***	***JoSAE***	***sae***	***saery***	***saeRobust***	***msae***	***hbsae***	***BayesSAE***	***mme***	***emdi***
Standard variance estimation	√		√				√			√
Adjusted variance estimation										√
Unknown sampling variances	√									
Heteroscedasticity		√								
Spatial correlation			√							√
Spatio-temporal correlation			√							
Temporal correlation				√						
Robust					√					√
Robust, spatial correlation					√					√
Robust, (spatio-)temporal correlation					√					
Multivariate						√				
Bayesian formulation							√	√		
Gaussian multinomial									√	
Transformation (log, arcsin)										√
Measurement error										√

Table 3.1: Overview of implemented area-level models in R packages available on CRAN.

the results to Excel and OpenDocument Spreadsheet are provided.

- Plus a stepwise variable selection algorithm for area-level models is included in **emdi** to allow the user to build a model based on information criteria.

Thus, the newly introduced package version 2.0.1 extends the current version 1.1.7 by various area-level models, but stays in line with the user-friendly orientation of the existing version. The structure of the paper can be described as follows. Section 3.2 introduces the statistical methods implemented in the package. The included example data sets are presented in Section 3.3. Section 3.4 provides an illustrative description of the functions using the example data sets. While Section 3.4.1 guides the reader from model building to model diagnostics of a standard FH model and exporting the results to Excel, Section 3.4.2 follows with relatively short descriptions of how to build the different extended area-level models. Finally, Section 3.5 concludes and gives an outlook.

3.2 Statistical methodology

Area-level models for the estimation of indicators like means, totals or shares have been added to the new package release (2.0.1). These comprise the area-level model by Fay and Herriot (1979) and several extensions of this standard model to account for issues that may come up in real data applications. To measure the precision of those models, respective MSE estimators have been integrated following the literature.

3.2.1 Standard Fay-Herriot model

Throughout the paper, a finite population U is assumed that consists of N units that are subdivided into D domains or areas of specific sizes $N_1, ..., N_D$. Then a random sample of size n can be drawn from U and partitioned into D areas with $n_1, ..., n_D$ observations per domain.
The FH model links area-level direct estimators that are based on survey data to covariates aggregated on an area level that stem from e.g., administrative (like register or census) data or alternative data sources (like satellite, social media or mobile phone data). The FH model is composed of two levels. The first one is the sampling model

$$\hat{\theta}_i^{\text{Dir}} = \theta_i + e_i, \quad i = 1, \ldots, D.$$

$\hat{\theta}_i^{\text{Dir}}$ is an unbiased direct estimator for a population indicator of interest θ_i, for instance a mean or a ratio. e_i stands for independent and normally distributed sampling errors with $e_i \overset{ind}{\sim} N(0, \sigma_{e_i}^2)$. Even though the model assumes known sampling variances, in practical applications $\sigma_{e_i}^2$ are usually unknown and have to be estimated from the unit-level sample data (Rivest and Vandal, 2003; Wang and Fuller, 2003; You and Chapman, 2006). Package **emdi** provides a non-parametric bootstrap for estimating the variances of the direct estimator (Alfons and Templ, 2013). To allow for complex survey designs, sampling weights (w) can be considered in the direct estimation (Horvitz and Thompson, 1952). For example, an estimator for the population mean θ_i of a continous variable of interest y for each area i is estimated by

$$\hat{\theta}_i^{\text{Dir}} = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

where the index j indicates an individual with $j = 1, ..., n_i$ in the i-th area. The second level links the target indicator θ_i linearly to area-specific covariates $\boldsymbol{x}_i$,

$$\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i,$$

where $\boldsymbol{\beta}$ is a vector of unknown fixed-effect parameters, u_i is an independent and identically normally distributed random effect with $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$.
The combination of the sampling and the linking model leads to a special linear mixed model

$$\hat{\theta}_i^{\text{Dir}} = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i + e_i, \quad i = 1, \ldots, D. \tag{3.1}$$

The empirical best linear unbiased estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ are computed by weighted least square theory. The empirical best linear unbiased predictor (EBLUP) of θ_i is obtained by substituting the variance parameter σ_u^2 with an estimate. The resulting estimator can then be written as

$$\begin{aligned} \hat{\theta}_i^{\text{FH}} &= \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} + \hat{u}_i \\ &= \hat{\gamma}_i \hat{\theta}_i^{\text{Dir}} + (1 - \hat{\gamma}_i) \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}. \end{aligned} \tag{3.2}$$

The EBLUP/FH estimator can be understood as a weighted average of the direct estimator $\hat{\theta}_i^{\text{Dir}}$ and a regression-synthetic part $\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$. The estimated shrinkage factor $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}$ puts more weight on the direct estimator when the sampling variance is small and vice versa. Areas for which no direct estimation results exist because the sample size is zero or the results may not be published are called out-of-sample domains. For those domains the prediction reduces to the regression-synthetic component $\hat{\theta}_{i,\text{out}}^{\text{FH}} = \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}$ (Rao and Molina, 2015).

Estimation methods for $\boldsymbol{\sigma_u^2}$

The variance of the random effects has to be estimated. Commonly used approaches are the FH method-of-moment estimator (Fay and Herriot, 1979), the ML, and the REML estimators (Rao and Molina, 2015). The likelihood methods are known to perform more efficiently than the methods of moments (Rao and Molina, 2015). The commonly used methods can produce negative variance estimates that are supposed to be strictly positive. In the estimation methods mentioned above, negative variance estimates are set to zero ($\hat{\sigma}_u^2 = max(\tilde{\sigma}_u^2, 0)$) resulting in zero estimates of the shrinkage factor γ_i. Therefore no weight is put on the direct estimator ignoring its possible reliability. This poses a problem especially when the number of areas is small. To avoid this so-called over-shrinkage problem, Li and Lahiri (2010) and Yoshimori and Lahiri (2014) proposed methods that adjust the respective likelihoods of the standard ML and REML approaches by a factor:

$$L_{\text{adj}}(\sigma_u^2) = A \times L\left(\sigma_u^2\right),$$

where A denotes the adjustment factor and $L(\sigma_u^2)$ the given likelihood function. The proposed adjustment factors are:

- by Li and Lahiri (2010): $A = \sigma_u^2$,
- by Yoshimori and Lahiri (2014): $A = \left(tan^{-1}\left(\sum_{i=1}^{D} \gamma_i\right)\right)^{1/D}$.

Simulation studies conducted by Yoshimori and Lahiri (2014) showed that the adjusted Yoshimori-Lahiri methods are preferable when the variance of the random effect is small relative to the sampling variance. Otherwise the adjusted Li-Lahiri methods are recommended. Package **emdi** offers six different variance estimation methods: standard ML (`ml`) and REML (`reml`), adjusted ML and REML following Li and Lahiri (2010) (`amrl`, `ampl`) and Yoshimori and Lahiri (2014) (`amrl_yl`, `ampl_yl`).

3.2.2 Extended area-level models

In real data applications problems might occur that were theoretically not expected or assumptions of the standard FH model, e.g., normality and independency of the error terms, may be violated. The following Section outlines the extensions of the standard FH model that are implemented in package **emdi**.

Transformations

When working with right skewed data like income, wealth or business data, the assumptions of a linear relation between the response and the explanatory variables and normality of both error terms (u_i and e_i) of the FH model may be violated. Applying a log-transformation could be a reasonable solution to meet these model assumptions (Neves et al., 2013; Kreutzmann, Marek, et al., 2019). In package **emdi**, the direct estimates and their variances are transformed following Neves et al. (2013):

$$\hat{\theta}_i^{\text{Dir*log}} = \log\left(\hat{\theta}_i^{\text{Dir}}\right),$$
$$\mathsf{VAR}(\hat{\theta}_i^{\text{Dir*log}}) = \left(\hat{\theta}_i^{\text{Dir}}\right)^{-2} \mathsf{VAR}\left(\hat{\theta}_i^{\text{Dir}}\right),$$

where the *log notation stands for the logarithmic transformed scale. To obtain the FH estimator on the transformed scale $\hat{\theta}_i^{\text{FH*log}}$, $\hat{\theta}_i^{\text{Dir}}$ is substituted by $\hat{\theta}_i^{\text{Dir*log}}$ and $\mathsf{VAR}(\hat{\theta}_i^{\text{Dir*log}})$ serves as estimate for the sampling variances ($\sigma_{e_i}^2$) in Equation 3.2. Since the logarithm is a nonlinear transformation, the final FH estimates on the original scale require a bias correction after the back-transformation (Slud and Maiti, 2006; Sugawasa and Kubokawa, 2017). Package **emdi** allows to choose two options:

1. A "crude" method (`bc_crude`) that takes the properties of the log-normal distribution into account (Rao, 2003; Neves et al., 2013):

$$\hat{\theta}_i^{\text{FH, crude}} = exp\left\{\hat{\theta}_i^{\text{FH*log}} + 0.5\text{MSE}\left(\hat{\theta}_i^{\text{FH*log}}\right)\right\}.$$

2. A bias correction suggested by Slud and Maiti (2006) (`bc_sm`) that further regards the bias due to the random effects:

$$\hat{\theta}_i^{\text{FH, Slud-Maiti}} = exp\left\{\hat{\theta}_i^{\text{FH*log}} + 0.5\hat{\sigma}_u^2\left(1 - \hat{\gamma}_i^{\text{*log}}\right)\right\}.$$

The FH estimator on the transformed scale is denoted by $\hat{\theta}_i^{\text{FH*log}}$ and accordingly $\text{MSE}(\hat{\theta}_i^{\text{FH*log}})$ stands for a MSE estimator on the transformed scale, e.g., the Prasad-Rao or Datta-Lahiri MSE (cf. Section 3.2.3). The Slud-Maiti back-transformation is derived for the ML variance estimation of the random effect and cannot be applied in the presence of out-of-sample domains, because the back-transformation contains the estimate of the shrinkage factor on domain level. In those cases, the "crude" method can be applied which allows to use also other variance estimation methods.

Another transformation provided by package **emdi** is the arcsin transformation that is widely used when the direct estimator of the FH model is a ratio (Casas-Cordero et al., 2016; Schmid, Bruckschen, et al., 2017). Package **emdi** automatically transforms the direct estimates and the sampling variances as suggested by Jiang, Lahiri, Wan, and Wu (2001):

$$\hat{\theta}_i^{\text{Dir*arcsin}},$$

where the *arcsin denotes the arcsin transformed scale and $\tilde{n}_i$ the effective sample size which can be described as the sample size adjusted by the sampling design (Jiang, Lahiri, Wan, and Wu, 2001). The FH model is estimated using Equation 3.2 and the results are additionally truncated to the interval $[0, \pi/2]$ to ensure results between 0 and 1, if needed. To obtain final estimates on the original scale, the final estimation results must be subjected to a back-transformation. Two different back-transformations are available in **emdi**:

1. A "naive" back-transformation (`naive`):

$$\hat{\theta}_i^{\text{FH, naive}} = sin^2\left(\hat{\theta}_i^{\text{FH*arcsin}}\right).$$

2. A back-transformation with bias-correction (`bc`) following Sugawasa and Kubokawa (2017) and Hadam et al. (2020):

$$\hat{\theta}_i^{\text{FH, bc}} = \int_{-\infty}^{\infty} sin^2(t) \frac{1}{2\pi \frac{\hat{\sigma}_u^2 \sigma_{e_i}^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}} exp\left(-\frac{\left(t - \hat{\theta}_i^{\text{FH*arcsin}}\right)^2}{2\frac{\hat{\sigma}_u^2 \sigma_{e_i}^2}{\hat{\sigma}_u^2 + \sigma_{e_i}^2}}\right) dt.$$

Spatial FH model

The standard FH model assumes independency of the random effects. When working with geographical areas, assuming correlated random effects to incorporate a certain neighbouring structure can be valuable. Package **emdi** contains the spatial FH model introduced by Petrucci and Salvati (2006) that considers a simultaneously autoregressive process of order one, SAR(1). Compared to the standard model, the estimation differs mainly by discarding the assumptions of independent random effects and estimating a spatial autoregressive coefficient (ρ) which takes values between -1 and 1. The higher the absolute value, the stronger the relationship with the neighboring areas. The random effect u_i in Equation 3.1 is replaced by

$$\boldsymbol{u} = \rho_1 \boldsymbol{W}\boldsymbol{u} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0}_D, \sigma_1^2 \boldsymbol{I}_D), \tag{3.3}$$

with $\boldsymbol{W}$ being the $D \times D$ row standardized proximity matrix that describes the neighbourhood structure of the areas, $\boldsymbol{0}_D$ a vector of zeros and $\boldsymbol{I}_D$ the $D \times D$ identity matrix. The random effects $\boldsymbol{u}$ of Equation 3.3 follow a SAR(1). When normality of the random effects is assumed, the model can be fitted by ML (`ml`) and REML (`reml`). The application of spatial FH models should be considered when no geographic auxiliary variables are available to capture the spatial relation or when ρ_1 is larger than 0.5 (Bertarelli et al., 2019). Even before estimating the

model, package **emdi** enables the testing for spatial correlation by the Moran's I and Geary's C statistics (Cliff and Ord, 1981; Pratesi and Salvati, 2008). While Moran's I mimics an usual correlation coefficient whose values range from -1 and 1, Geary's C takes values between 0 and 2 (0: positive, 1: no, 2: negative spatial autocorrelation). Both statistics behave inversely to each other.

Robust area-level models

For the case of influential outlying observations, package **emdi** allows for robust versions of the standard and the spatial FH model. The theory is extensively studied in Warnholz (2016) that extended the robust estimation procedure for linear mixed models suggested by Sinha and Rao (2009) to area-level models. The model fitting can be understood as a robustified ML version that also contains an influence function together with a tuning constant `k`. The recommendation is to set the tuning constant to 1.345 (Sinha and Rao, 2009). When non-symmetric outliers are expected to influence the robust estimation, a bias correction should be involved. This correction can be controlled by a multiplyer constant `c` that is used for the bias correction. For further details we also refer to Chambers, Chandra, et al. (2014) and Schmid, Tzavidis, et al. (2016).

Measurement error model

The standard FH model is based on the assumption that the covariates are measured without error (Fay and Herriot, 1979). This characteristic is typically assumed because census or register data are used as auxiliary information. However, when the covariate information stems from larger surveys or alternative data sources this assumption can be violated. Package **emdi** includes an implementation of the measurement error (ME) model developed by Ybarra and Lohr (2008). To account for the ME in the covariates $\boldsymbol{x}_i$, they modified the shrinkage factor as follows:

$$\gamma_i = \frac{\sigma_u^2 + \boldsymbol{\beta}^\top \boldsymbol{C}_i \boldsymbol{\beta}}{\sigma_u^2 + \boldsymbol{\beta}^\top \boldsymbol{C}_i \boldsymbol{\beta} + \sigma_{e_i}^2},$$

where the $\boldsymbol{C}_i$ stands for variance-covariance matrix of the covariates which needs to be given to the model. The modified shrinkage factor pulls more weight on the direct estimator when the variances of the covariates are large. For the estimation of the $\boldsymbol{\beta}$s and the σ_u^2, they used a modified method of weighted least squares and a moment estimator, respectively. Additional details are available in Ybarra and Lohr (2008).

3.2.3 Mean squared error estimation

To evaluate the accuracy of the EBLUP estimates, the MSE is the most common measure used in SAE (Rao and Molina, 2015). Package **emdi** offers a variety of MSE estimators stemming from both analytical determination and resampling strategies like boostrap and jackknife methods. Table 3.2 gives an overview about the included MSE approaches. For each area-level model presented in Section 3.2.1 and 3.2.2 the provided MSE type(s) is (are) shown. Please

Model	Type of MSE	Reference
Standard FH (depending on variance estimation of σ_u^2)		
`ml`/`ampl_yl`	Analytical	Datta and Lahiri (2000)
`reml`/`amrl_yl`	Analytical	Prasad and Rao (1990)
`ampl`/`amrl`	Analytical	Li and Lahiri (2010)
`ml`/`reml` (out-of-sample)	Analytical	Rao and Molina (2015)
Transformations		
`log` (depending on back-transformation)		
`bc_crude`	Analytical	Rao (2003), Neves et al. (2013)
`bc_sm`	Analytical	Slud and Maiti (2006)
`arcsin` (depending on back-transformation)		
`naive`	Jackknife	Jiang, Lahiri, Wan, and Wu (2001)
	Weighted Jackknife	Jiang, Lahiri, Wan, and Wu (2001) and Chen and Lahiri (2002)
	Parametric bootstrap	Hadam et al. (2020)
`bc`	Parametric bootstrap	Hadam et al. (2020)
Spatial FH (depending on variance estimation)		
`ml`/`reml`	Analytical	Singh et al. (2005)
`ml`/`reml`	Parametric bootstrap	Molina, Salvati, et al. (2009)
`reml`	Nonparametric bootstrap	Molina, Salvati, et al. (2009)
Robust FH		
	Pseudolinear	Warnholz (2016)
	Parametric bootstrap	Warnholz (2016)
FH with ME		
	Jackknife	Jiang, Lahiri, and Wan (2002)

Table 3.2: Overview of the MSE estimation options of the `fh` function.

refer to the quoted references for extensive formulas and derivations. As additional measure of variability of the direct and FH estimates, within various functions and methods of package **emdi**, the coefficient of variation (CV) is provided: $CV = \sqrt{\widehat{MSE}(\hat{\theta}_i)}/\hat{\theta}_i$, where $\hat{\theta}_i$ either stands for $\hat{\theta}_i^{\text{Dir}}$ or $\hat{\theta}_i^{\text{FH}}$.

3.3 Data sets

The version 1.1.7 of package **emdi** contains a sample (`eusilcA_smp`) and a population data set (`eusilcA_pop`) at a household level.

The data generating processs for both data sets is extensively described in Kreutzmann, Pannier, et al., 2019. Besides the modification of not producing out-of-sample domains for the area-level version of the data sets, the process is almost equivalent. As basis for the data

Variable	Meaning
Sample data set	
Domain	Austrian districts
Mean	Mean of the equivalized household income
MTMED	Share of households who earn more than the national median income
Cash	Mean employee cash or near cash income
Var_Mean	Variance of equivalized household income
Var_MTMED	Variance of share of households who earn more than the national median income
Var_Cash	Variance of employee cash or near cash income
n	Effective sample sizes
Population data set	
Domain	Austrian districts
eqsize	Equivalized household size according to the modified OECD scale
cash	Employee cash or near cash income
self_empl	Cash benefits or losses from self-employment (net)
unempl_ben	Unemployment benefits (net)
age_ben	Old-age benefits (net)
surv_ben	Survivor's benefits (net)
sick_ben	Sickness benefits (net)
dis_ben	Disability benefits (net)
rent	Income from rental of a property or land (net)
fam_allow	Family/children related allowances (net)
house_allow	Housing allowances (net)
cap_inv	Interest, dividends, profit from capital investments in unincorporated business (net)
tax_adj	Repayments/receipts for tax adjustment (net)
ratio_n	Ratios of the population size per area and the total population size

Table 3.3: Variables of the aggregated data sets. The Domain variables are factors, the rest of the variables are numeric. Except for the variables Domain and ratio_n, the observations of all variables of the population data set consist of the mean values per district.

sets serves the synthetic Austrian European Union Statistics on Income and Living Conditions (EU-SILC) data set (eusilcP) from 2006 of the **simFrame** package (Alfons, Templ, and Filzmoser, 2010). The lowest regional level in the eusilcP data set consists of the nine Austrian states. Based on certain population size and income criteria, households were allocated to 94 Austrian districts resulting in the synthetic population data set eusilcA_pop. For the eusilcA_smp data set, a sample was drawn following a stratified random sampling process using the districts as strata. To show the usage of the FH model and its extensions, area-level data is required. The area-level survey and population data sets, eusilcA_smpAgg and eusilcA_popAgg, are obtained by aggregation on the district level with the help of the direct function of the package **emdi**. The direct estimates in eusilcA_smpAgg are the weighted mean equivalized household income Mean, the ratio of households that earn more than the national median income (MTMED) and their variances. These are based on the equiv-

alized household income `eqIncome` in `eusilcA_smp` corresponding to the total income of a household divided by the size of the household that is equalised by the modified equivalence scale of the Organisation for Economic Co-operation and Development (OECD) (Hagenaars et al., 1994). Additionally, the mean of the variable `cash`, its variance and the sample sizes are included in `eusilcA_smpAgg` since these are used in the model extensions. The population data set `eusilcA_popAgg` contains a variety of variables that describe different income sources of households and a variable that describes the ratios of the population sizes per area and the total population size `ratio_n`. The variable `Domain` exists in both data sets and identifies the different districts. Both data sets have 94 observations standing for the 94 Austrian districts, the sample data set `eusilcA_smpAgg` contains eight variables and the population data set `eusilcA_popAgg` 15. Table 3.3 provides an overview of all included variables of the sample and population data set. For the creation of the proximity matrix used in the spatial FH model and also for the usage of the `map_plot` function, a shape file is needed. A shape file `shape_austria_dis` (`.rda` format, `SpatialPolygonsDataFrame`) for the 94 districts of Austria is provided. It stems from the SynerGIS website (Bundesamt für Eich- und Vermessungswesen, 2017). The data set `eusilcA_prox` comprising an exemplary proximity matrix is also added to package **emdi**. The creation of `eusilcA_prox` is described in Section 3.4.1.

3.4 Functionality and case studies

While the theoretical background of the implemented area-level models has been introduced in Section 3.2, the focus of Section 3.4 lies on the functionality and the work flow in R. All of the contained area-level models can be applied by one function: `fh`. Table 3.4 gives an overview of the 20 input arguments of function `fh`, together with a short description and default settings if specified. Not every argument needs a specification for every estimated model. Depending on the area-level model, different arguments have to be determined (see Table 3.6 in Appendix 3.6). The flow diagram of Figure 3.1 demonstrates the estimation possibilities of a standard FH model introduced in Section 3.2.1.

In line with the `direct` and `ebp` functions of package version 1.1.7, the S3 object system is used for function `fh` (Chambers and Hastie, 1992). All three return objects of class `emdi`, but in addition, the application of function `direct` leads to a `direct` object, and of functions `ebp` and `fh` to objects of class `model`. The latter two are further classified into `ebp` and `fh` objects. Even though all of the returned objects contain ten components, not every component is available for each estimation method such that in these cases they are indicated as `NULL` (see Table 3.5). Furthermore, the `model` component differs for the two model classes. The components for the objects of class `fh` are provided in Table 3.7 in Appendix 3.7. Not all of the components are available for every area-level model, e.g., the shrinkage factors per domain are not provided for the spatial and robust model extensions as they do not enable an intuitive

Argument	Description	Default
fixed	Formula of fixed-effects part of linear mixed model	
vardir	Domain-specific sampling variances of the direct estimates	
combined_data	Combined sample and census data set	
domains	Domain indentifier for combined_data	NULL
method	Model fitting method	reml
interval	Lower and upper limit for the variance estimation	NULL
k	Tuning constant for robust estimation	1.345
c	Bias correction multiplyer constant for robust estimation	1
transformation	Type of transformation	no
backtransformation	Type of back-transformation	NULL
eff_smpsize	Effective sample sizes for the arcsin transformation	NULL
correlation	Correlation of random effects	no
corMatrix	Proximity matrix for the spatial model	NULL
Ci	Array of the variance-covariance matrix of the explanatory variables for each area for the ME model	NULL
tol	Tolerance value for the variance estimation	0.0001
maxit	Maximum number of iteration for the variance estimation	100
MSE	MSE estimation	FALSE
mse_type	Type of MSE estimator	analytical
B	Number of bootstrap iteration for estimating a bootstrap MSE	50
seed	Seed for random number generator	123

Table 3.4: Input arguments of function fh.

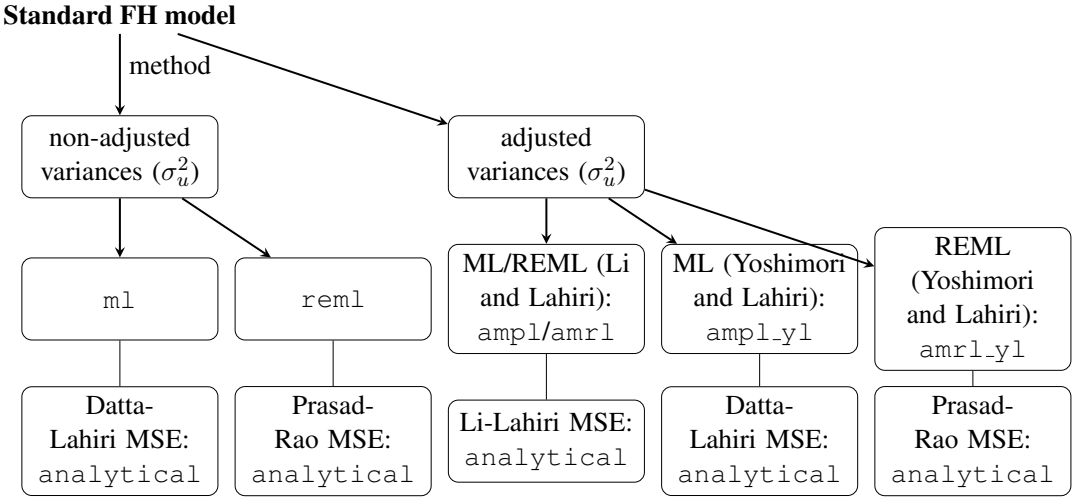


Figure 3.1: Overview of the standard FH model and adjusted variance estimation methods.

	Name	Description	Available for		
			direct	model ebp	model fh
1	ind	Point estimates per area	√	√	√
2	MSE	Variance/MSE estimaties per area	√	√	√
3	transform_param	Transformation and shift parameters		√	
4	model	Fitted model		√	√
5	framework	List for data description	√	√	√
6	transformation	Type of transformation		√	√
7	method	Estimation method		√	√
8	fixed	Formula of fixed effects		√	√
9	call	Function call	√	√	√
10	successful_bootstraps	Number of successful bootstraps	√		√

Table 3.5: The ten emdi object components distinguished in direct, ebp and fh. More detailed information are provided by the package documentation.

interpretation. Due to the consistent structure, all functions and methods of **emdi** version 1.1.7 can be applied to objects of class fh. Additionally, new functions and methods are available for the area-level models. Figure 3.2 demonstrates the steps of a full data analysis procedure and the respective functions from model building and diagnostics to presenting the results. Section 3.4.1 explains the procedure shown in Figure 3.2 step by step for the standard FH model by using the Austrian EU-SILC data described in Section 3.3. To understand how the different extended area-level models are fitted with function fh, Section 3.4.2 shortly gives instructions.

3.4.1 Estimation procedure for the standard Fay-Herriot model

The aim of the illustrative example is to estimate the equivalized income for the 94 Austrian districts. The package and the example data sets are loaded as follows:

```
R> library("emdi")
R> data("eusilcA_popAgg")
R> data("eusilcA_smpAgg")
```

Combine input data

The function fh requires one data set (argument combined_data) that comprises the sample and population data. Thus, the data set has to contain all variables of the formula object fixed, the variances of the direct estimates and optionally, a domain identifier. In case the sample

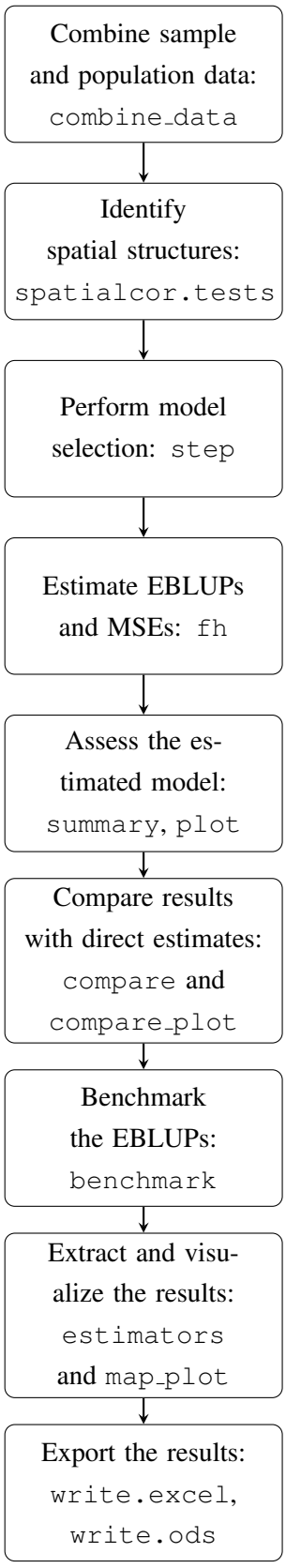


Figure 3.2: Estimation procedure for area-level models.

and population data are only available separately, a merging function `combine_data` is provided. The necessary arguments are both data sets and characters specifying the domain indicator for the respective data sets.

```
R> combined_data <- combine_data(
+    pop_data = eusilcA_popAgg,
+    pop_domains = "Domain",
+    smp_data = eusilcA_smpAgg,
+    smp_domains = "Domain")
```

Identify spatial structures

With the help of a proximity matrix, the Moran's I and Geary's C test statistics can be computed to identify spatial structures by the `spatialcor.tests` command. For the creation of the proximity matrix, the shapefile has to be loaded. We load the Austrian shapefile that is provided by package **emdi** for our example and merge it to the sample data set by using the respective domain identifiers with the help of the `merge` method from package **sp** (Pebesma and Bivand, 2005). Before merging, we sort the Austrian shapefile corresponding to the order of the domains in the sample data.

```
R> library("sp")
R> load_shapeaustria()
R> shape_austria_dis <- shape_austria_dis[
+    order(shape_austria_dis$PB),]
R> austria_shape <- merge(shape_austria_dis,
+    eusilcA_smpAgg, by.x = "PB",
+    by.y = "Domain", all.x = F)
```

Then the `poly2nb` and `nb2mat` functions of the **spdep** package (Bivand and Wong, 2018) are used. While `poly2nb` generates a list of neighbours that share joint boundaries, `nb2mat` computes a weights matrix. The `style` argument has to be set to `W`, as a row standardized proximity matrix is required.

```
R> library("spdep")
R> rel <- poly2nb(austria_shape,
+    row.names = austria_shape$PB)
R> eusilcA_prox <- nb2mat(rel,
+    style = "W", zero.policy = TRUE)
```

Thus, a row standardized proximity matrix is generated that initially had weights amounting to one if an area shares a boundary with another area and to zero when the respective areas are not neighbours. Function `spatialcor.tests` makes use of the `moran.test` and `geary.test` functions with their respective default settings of package **spdep**. The input arguments are the created matrix and the direct estimates.

```
R> spatialcor.tests(
+    direct = combined_data$Mean,
+    corMatrix = eusilcA_prox)
```

```
  Statistics     Value      p.value
1  Moran's I 0.2453677 5.607958e-05
2  Geary's C 0.6238681 2.473294e-03
```

Since the output indicates only a weak positive spatial autocorrelation, the following estimation procedure does not consider the integration of a correlation structure of the random effects.

Perform model selection

Besides theoretical considerations on which auxiliary variables should be part of the model, the decision for the best model should be based on information criteria like the Akaike or Bayesian information criterion (AIC, BIC). Many applications use selection techniques based on linear regression (Casas-Cordero et al., 2016; Schmid, Bruckschen, et al., 2017). Instead, package **emdi** provides the AIC, BIC, the Kullback information criterion (KIC) and their bootstrap and bias corrected versions (AICc, AICb1, AICb2, KICc, KICb1, KICb2) especially developed for FH models by Marhuenda, Morales, et al. (2014). These criteria are also included in the package **sae**, but package **emdi** enables a stepwise variable selection procedure based on the chosen information criteria comparable to the `step` function for `lm` models of package **stats** (R Core Team, 2020). The most important input arguments are an object of class `fh` and the `direction` of the stepwise search ("`both`", "`backward`", "`forward`"). In this example, the default setting "`backward`" and the "`KICb2`" information criterion is used. In the `fixed` argument of the `fh` function, the variables equivalized household size (`eqsize`), employee cash (`cash`), cash benefits from self-employment (`self_empl`) and unemployment benefits (`unempl_ben`) are included. For a valid comparison of models based on information criteria the model fitting `method` has to be `ml`. The output shows the stepwise removal of variables until the lowest KICb2 is reached, the function call and an overview of the estimated coefficients of the final recommended model.

```
R> fh_std <- fh(fixed = Mean ~ cash + self_empl + unempl_ben,
+    vardir = "Var_Mean", combined_data = combined_data,
+    domains = "Domain", method = "ml")
R> step(fh_std, criteria = "KICb2")

Start: KICb2 = 1709.42
Mean ~ cash + self_empl + unempl_ben

             df  KICb2
- unempl_ben  1 1708.3
<none>          1709.4
- self_empl   1 1763.0
- cash        1 1808.6

Step: KICb2 = 1708.33
Mean ~ cash + self_empl

            df  KICb2
<none>         1708.3
- self_empl  1 1765.3
- cash       1 1816.1


Call:
fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean",
    combined_data = combined_data,
    domains = "Domain", method = "ml", MSE = FALSE)

Coefficients:
             coefficients    std.error   t.value      p.value
(Intercept)  3070.512311 635.94290168  4.828283 1.377153e-06
cash            1.059385   0.07049025 15.028815 4.754350e-51
self_empl       1.745636   0.22017394  7.928443 2.219112e-15
```

KICb2 is the lowest when the variable `unempl_ben` is removed. Therefore, the formula `Mean cash + self_empl` is used in the following.

Estimate EBLUPs and MSEs

The standard FH model is built. In addition to the `fixed` part, required arguments are `vardir` and `combined_data`. We specify the `domains` (if the `domains` argument is set to `NULL`, the domains are numbered consecutively) and activate the MSE estimation.

```
R> fh_std <- fh(fixed = Mean ~ cash + self_empl,
+     vardir = "Var_Mean", combined_data = combined_data,
+     domains = "Domain", method = "ml", MSE = TRUE)
```

Assess the estimated model

In many publications using FH models, model diagnostics are not or only little discussed. One reason for this might be the lack of existing implementation of those measures in R or other statistical software. The `summary` method of **emdi** provides additional information about the data and model components, in particular the chosen estimation methods, the number of domains, the log-likelihood, the information criteria by Marhuenda, Morales, et al. (2014), the R^2 and the adjusted R^2 proposed by Lahiri and Suntornchost (2015). Additionally, measures to validate model assumptions about the standardized realized residuals and the random effects are provided: skewness and kurtosis (`skewness` and `kurtosis` of package **moments**, Komsta and Novomestky, 2015) of the standardized realized residuals and the random effects and the test statistics with corresponding p value of the Shapiro-Wilks-test for normality of both error terms. As the introduced area-level models do not assume a homoscedastic sampling distribution, for the `summary` and `plot` methods the realized residuals ($\hat{e}_i$) are standardized by: $\hat{e}_i^{\text{std}} = \hat{e}_i / \sigma_{e_i}$. The `summary` output differs slightly for the different implemented area-level models. For example, log-likehoods and thus information criteria are not available in theory for the robust and the ME model.

```
R> summary(fh_std)
```

```
Call:
fh(fixed = Mean ~ cash + self_empl, vardir = "Var_Mean",
    combined_data = combined_data,
    domains = "Domain", method = "ml", MSE = TRUE)

Out-of-sample domains:  0
In-sample domains:  94

Variance and MSE estimation:
Variance estimation method:  ml
Estimated variance component(s):  1371195
MSE method:  datta-lahiri
```

```
Coefficients:
            coefficients    std.error   t.value      p.value
(Intercept)  3070.512311 635.94290168  4.828283 1.377153e-06
cash            1.059385   0.07049025 15.028815 4.754350e-51
self_empl       1.745636   0.22017394  7.928443 2.219112e-15

Explanatory measures:
    loglike      AIC    AICc    AICb1    AICb2      BIC      KIC
1 -847.8303 1703.661 1703.91 1715.758 1703.461 1713.834 1707.661
      KICc    KICb1    KICb2        R2     AdjR2
1 1708.783 1720.632 1708.335 0.9212817 0.9482498

Residual diagnostics:
                         Skewness Kurtosis Shapiro_W Shapiro_p
Standardized_Residuals  0.3004662 3.971216 0.9840810 0.3119346
Random_effects         -0.4113238 3.086048 0.9839858 0.3072834

Transformation: No transformation
```

The output of the example shows that all domains have survey information and the variance of σ_u^2 amounts to 1371195. Further, all of the included auxiliary variables are significant even on a small significance level and their explanatory power is large with an adjusted R^2 of around 0.95. The results of the Shapiro-Wilk-test indicate that normality is not rejected for both errors. Graphical residual diagnostics are possible by the `plot` method.

```
R> plot(fh_std)
```

Figure 3.3 shows normal quantile-quantile (Q-Q) plots of the standardized realized residuals and random effects (Figure 3.3a) as well as plots of the kernel densities of the distribution of both error terms and for comparison a standard normal distribution (Figure 3.3b and 3.3c). Like in the **emdi** version 1.1.7, the user is free to modify the interface of the plots. The `label` and `color` arguments are easy to edit. Additionally, the overall appearance of the plots are changeable by the `gg_theme` argument as the plots are built with the **ggplot2** package (Wickham, 2016). We refer to the package documentation for a detailed description of how to customize the `plot` arguments. Figure 3.3 supports the results of the normality tests provided in the `summary` output, the distribution of the standardized random effects may be slightly skewed (Figure 3.3c). If one would not be satisfied with the results, applying a log-transformation could improve the distribution of the error terms.

Compare results with direct estimates

The FH results should be consistent with the direct estimates for domains with a small direct MSE and/or large sample sizes. Further, the precision of the direct estimates should be improved by using auxiliary information. The comparison of the direct and model-based (FH)

estimates can be done graphically by the generic function `compare_plot`. For the `fh` method the required input argument is an object of class `fh`. When the default settings of the command are used, the output consists of two plots: a scatter plot proposed by Brown et al. (2001) and a line plot. Besides the direct and FH estimates, the plot contains the fitted regression and the identity line. Both lines should not differ too much. Preferably, the model-based (FH) estimates should track the direct estimates within the line plot especially for domains with a large sample size/small MSE of the direct estimator. The points are ordered by decreasing MSE of the direct estimates. In addition, the input arguments `MSE` and `CV` can be set to `TRUE` leading to two extra plots, respectively. The MSE/CV estimates of the direct and model-based (FH) estimates are compared firstly via boxplots and secondly via ordered scatter plots (ordered by increasing CV of the direct estimates). Like for the `plot` command, a variety of customization options are offered, e.g., the label options (`label`), the format of the points (`shape`) and the style of the line (`line_type`).

```
R> compare_plot(fh_std, CV = TRUE, label = "no_title")
```

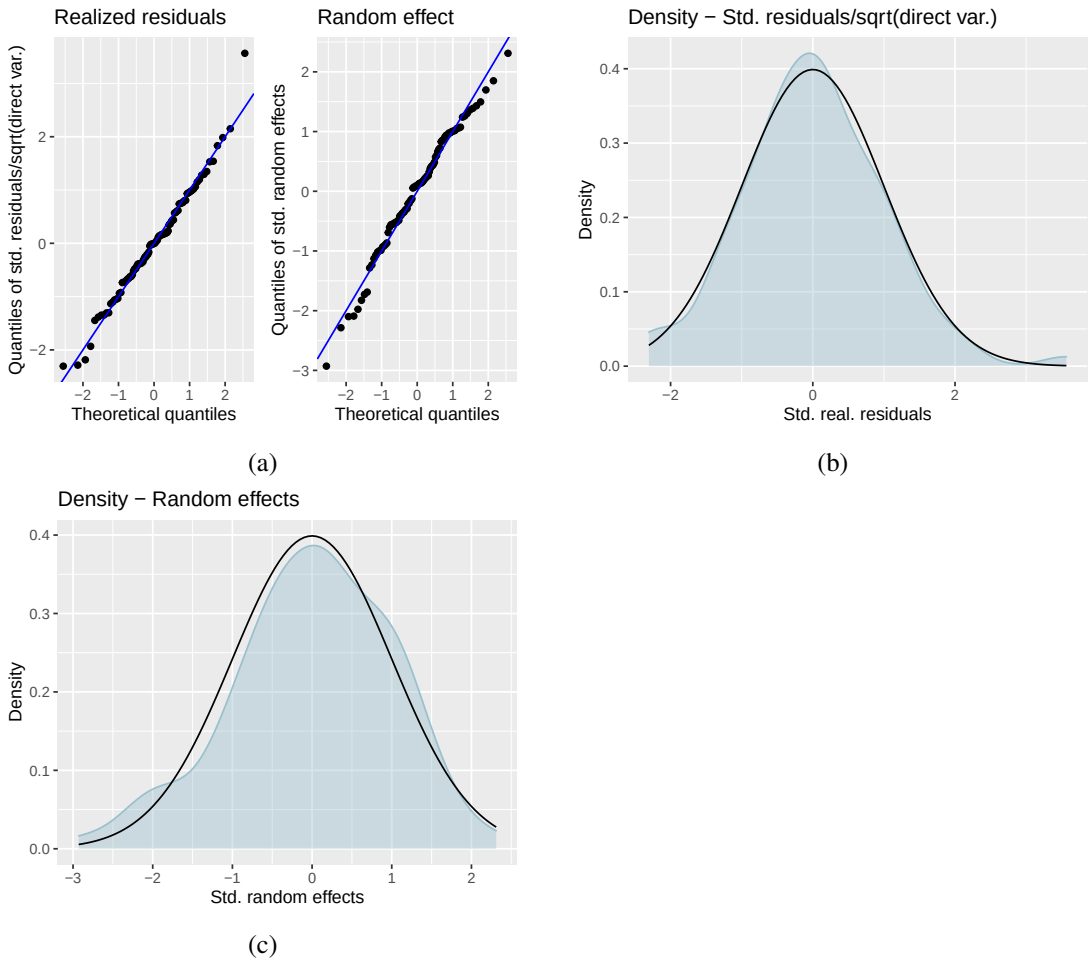


Figure 3.3: Output of `plot(fh_std)`: (a) normal quantile-quantile (Q-Q) plots of the standardized realized residuals and random effects, (b) and (c): kernel densities of the distribution of the standardized realized residuals and random effects (blue) in comparison to a standard normal distribution (black).

Except of one high value, the fitted regression and identity line of the scatter plot (Figure 3.4a) are relatively close. Note that the high value corresponds to the domain Eisenstadt (Stadt) with a very small sample size of 10 and the highest MSE of the direct estimates, so the direct estimator is very uncertain. Also the direct estimates are well tracked by the model-based (FH) estimates within the line plot (Figure 3.4b). The boxplot (Figure 3.4c) and the ordered scatter plot (Figure 3.4d) show that the precision of the direct estimates could be improved by the usage of the FH model in terms of CVs. Additionally, almost all of the CV values are less than 20% which is a common rule of the UK Office for National Statistics in order to determine whether estimation results should be published (Miltiadou, 2020).

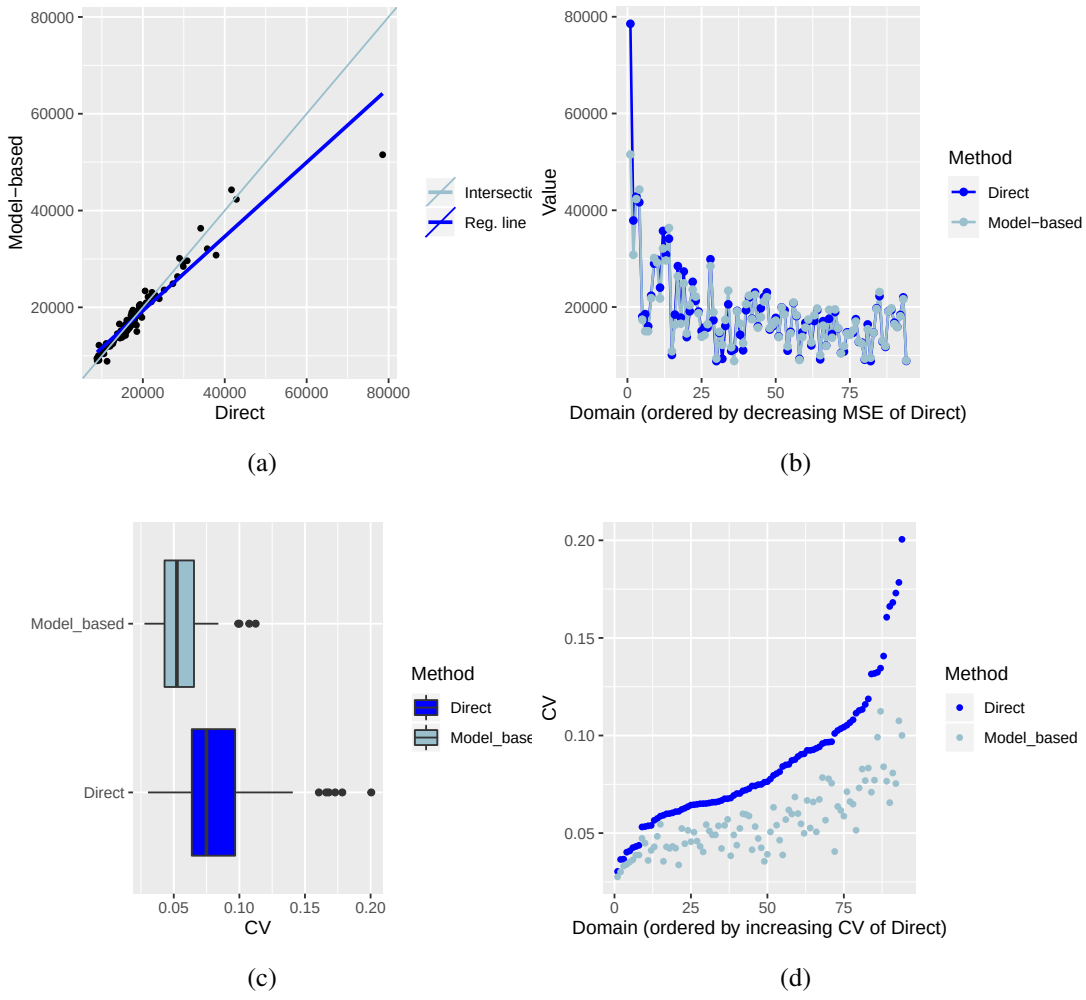


Figure 3.4: Output of `compare_plot(fh_std)`: (a) and (b) scatter and line plots of direct and model-based point estimates, (c) and (d) boxplot and scatter plots of the CV estimates of the direct and model-based (FH) estimates.

Further on, the function `compare` enables the user to compute a goodness of fit diagnostic (Brown et al., 2001) and a correlation coefficient of the direct estimates and the estimates of the regression-synthetic part of the FH model (Chandra et al., 2015). Following Brown et al. (2001) the difference between the model-based estimates and the direct estimates should not

be significant (null hypothesis). The Wald test statistic is specified as

$$W(\hat{\theta}_i^{\text{FH}}) = \sum_{i=1}^{D} \frac{\left(\hat{\theta}_i^{\text{Dir}} - \hat{\theta}_i^{\text{FH}}\right)^2}{\widehat{\text{VAR}}\left(\hat{\theta}_i^{\text{Dir}}\right) + \widehat{MSE}\left(\hat{\theta}_i^{\text{FH}}\right)}$$

and is χ^2-distributed with D degrees of freedom. When working with out-of-sample domains, those are not taken into account, because the direct estimates and their variances are missing. The input argument of function `compare` is an `fh` object.

```
R> compare(fh_std)

Brown test

Null hypothesis: EBLUP estimates do not differ significantly from the
      direct estimates

  W.value Df   p.value
 46.97181 94 0.9999874

Correlation between synthetic part and direct estimator:  0.94
```

The results of the goodness of fit statistic and the correlation coefficient confirm what the scatter and the line plot already indicated. In the example the null hypothesis is not rejected and the correlation coefficient indicates a strong positive correlation (0.94) between the direct and model-based (FH) estimates.

Benchmarking for consistent estimates

The idea of benchmarking is that the aggregated FH estimates should sum up to estimates of a higher regional level (τ):

$$\sum_{i=1}^{D} \xi_i \hat{\theta}_i^{\text{FH,bench}} = \tau,$$

where ξ_i stands for the share of the population size of each area in the total population size (N_i/N). In our example, the EBLUP estimates could get aggregated on a national level and then compared to or benchmarked with the Austrian mean equivalized income. Package **emdi** contains a benchmark function that allows the user to select three different options suggested by Datta, Ghosh, et al. (2011). A general estimator of the three options can be written as follows:

$$\hat{\theta}_i^{\text{FH,bench}} = \hat{\theta}_i^{\text{FH}} + \left(\sum_{i=1}^{D} \frac{\xi_i^2}{\phi_i}\right)^{-1} \left(\tau - \sum_{i=1}^{D} \xi_i \hat{\theta}_i^{\text{FH}}\right) \frac{\xi_i}{\phi_i}.$$

Depending on the weight ϕ_i, the formula leads to different benchmarking options. If ϕ_i equals ξ_i, all FH estimates are adjusted by the same factor (`raking`). A ratio adjustment (`ratio`) is being conducted if $\phi_i = \xi_i/\hat{\theta}_i^{\text{FH}}$. For the last option, $\hat{\theta}_i^{\text{FH}}$ is replaced by $\widehat{\text{MSE}}(\hat{\theta}_i^{\text{FH}})$ in the

ratio adjustment formula (`MSE_adj`). While the first option is a relatively naive approach, the latter two conduct larger adjustments for the areas with larger FH and MSE estimates, respectively. Thus, for the `benchmark` function the following arguments have to be specified: an object of class `fh`, a `benchmark` value, a vector containing the ξ_is (`share`) and the `type` of benchmarking. The output is a data frame with an extra column `FH_Bench` for the benchmarked EBLUP values. If the optional argument `overwrite` is set to `TRUE`, the benchmarked results are added to the `fh` object and the MSE estimates of the non benchmarked FH estimates are set to `NULL`. For the used example, the `benchmark` value is calculated by taking the mean of the variable `eqIncome` of the `eusilcA_smp` data frame. The ξ_is can be found in `eusilcA_popAgg` as `ratio_n`.

```
R> fh_bench <- benchmark(fh_std, benchmark = 20140.09,
+    share = eusilcA_popAgg$ratio_n, type = "ratio")
R> head(fh_bench)
```

```
               Domain   Direct       FH FH_Bench Out
1           Amstetten 14768.57 14242.04 14480.61   0
2               Baden 21995.72 21616.40 21978.49   0
3             Bludenz 12069.59 12680.38 12892.79   0
4      Braunau am Inn 10770.53 11925.82 12125.59   0
5             Bregenz 35731.20 32101.69 32639.43   0
6 Bruck-Mürzzuschlag 23027.37 22523.50 22900.79   0
```

It is recognizeable that for the first six Austrian districts the original estimates are slightly modified by the benchmarking.

Extract and visualize the results

With package **emdi** the user is able to produce data frames and high quality maps of the results to easily recognize geographic differences. Of course the `fh` object already outputs the components `ind` and `MSE` that provide the EBLUP and MSE estimates per area, respectively. The generic function `estimators` offers an easy way to overview the EBLUP, MSE and CV results of the direct estimates compared to the model-based (FH) results. The following output shows the EBLUP and MSE results for the first six domains in Austria.

```
R> head(estimators(fh_std, MSE = TRUE))
```

```
               Domain   Direct Direct_MSE       FH    FH_MSE
1           Amstetten 14768.57   926167.4 14242.04  599010.6
2               Baden 21995.72   446534.3 21616.40  356586.1
3             Bludenz 12069.59  1243265.0 12680.38  716040.1
4      Braunau am Inn 10770.53  1029502.4 11925.82  643500.2
5             Bregenz 35731.20  4467316.4 32101.69 1302156.0
6 Bruck-Mürzzuschlag 23027.37  1971664.0 22523.50  906339.2
```

While the highest equalized income of the considered domains was found in Bregenz, the lowest was estimated for Braunau am Inn. The MSE estimates of the EBLUPs are always lower than those of the direct estimates, indicating that the precision of the direct estimates could be improved with the help of the FH model.

Differences among the areas or hotspots of special interest are easier to detect on maps. With function `map_plot`, package **emdi** offers a user-friendly way to produce maps since creating maps can often become a time consuming task. The input arguments mainly consist of an object of class `emdi` and a spatial polygon of a shape file. The only issue that might come up is if domain identifiers in the data do not match to the respective identifiers of the shape file. In those cases, the input argument `map_tab` is required which is a data frame that contains the matching of the domain indentifiers of the population and the shape file data sets. For detailed instructions, we refer to Kreutzmann, Pannier, et al. (2019) and to the help page of function `map_plot`.

For producing maps of the 94 Austrian districts, the Austrian shape file has to be loaded. In addition to the `emdi` object, the `SpatialPolygonsDataFrame` object (`map_obj`) and a domain indicator (`map_dom_id`) have to be specified. The `map_tab` argument is not necessary since the identifiers match in our example. To allow for an easier comparison of the results, we adjust the scales of the maps using the `scale_points` argument.

```
R> load_shapeaustria()
R> map_plot(object = fh_std, MSE = TRUE,
+    map_obj = shape_austria_dis, map_dom_id = "PB",
+    scale_points = list(Direct = list(ind = c(8000, 60000),
+    MSE = c(200000, 10000000)), FH = list(ind = c(8000, 60000),
+    MSE = c(200000, 10000000))))
```

Figures 3.5a and 3.5c show the distribution of the estimated (direct vs. model-based) equivalized income across Austria. It is striking that white and light red tones dominate the map, indicating relatively low mean incomes of the districts. But in contrast, districts like for example Eisenstadt (Stadt), Urfahr-Umgebung and Mödling stand out having the largest incomes. Urfahr-Umgebung is also eye-catching when having a look at the MSE estimates (Figures 3.5b and 3.5d). The MSE of the direct and the FH estimates are quiet high. Probably a single wealthy household raised the mean income and also the variance. Figure 3.5b contains some districts with MSEs larger than the customized scaling (gray areas). Without the scaling it would have been hard to identify any differences in Figure 3.5d.

Export the results

Some users might have an interest to store the results separately or to use them for presentations. Excel provides many opportunities for that. Compared to some existing R packages, the **emdi** function `write.excel` does not only export the estimation results to Excel, but also the output of `summary`. The input arguments are again similar to the `estimators`

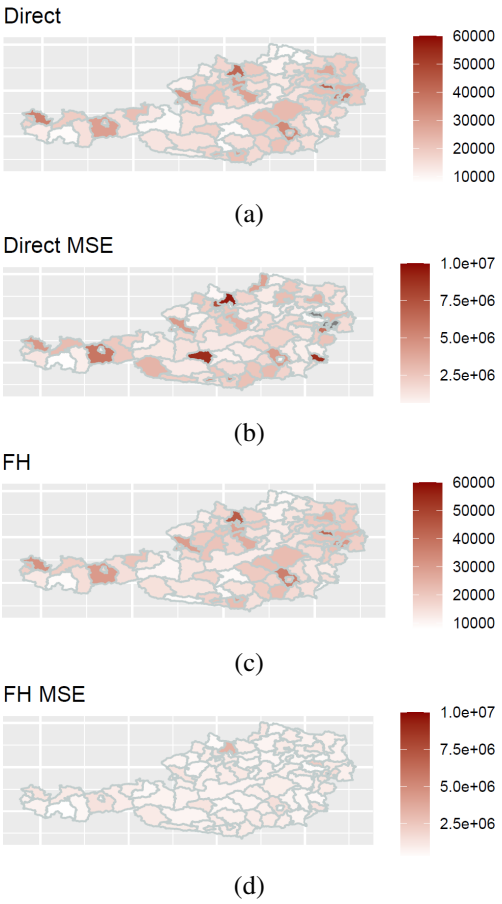


Figure 3.5: Output of `map_plot()`: Maps of the direct and FH estimates ((a) and (c)) with corresponding MSE estimates ((b) and (d)).

command except that the newly created path and filename of the spreadsheet `file` has to be specified. The output consists of a new Excel file which shows the `summary` output on the first sheet and the estimation results on the second sheet. The package **openxlsx** (Walker, 2018) has been used for the linkage with Excel. When working with Microsoft Windows an extra zipping applications for R is necessary for the usage of package **openxlsx** (Walker, 2018). Thus, the user is recommended to install RTools. For Linux and macOS zipping application are automatically installed. Using a similar syntax, the results can also be exported to OpenDocument Spreadsheets by the command `write.ods`. The difference to `write.excel` is that multipe files are created. The output of the FH model is exemplarily exported to Excel.

```
R> write.excel(fh_std,
+    file = "fh_std_output.xlsx",
+    MSE = TRUE, CV = TRUE)
```

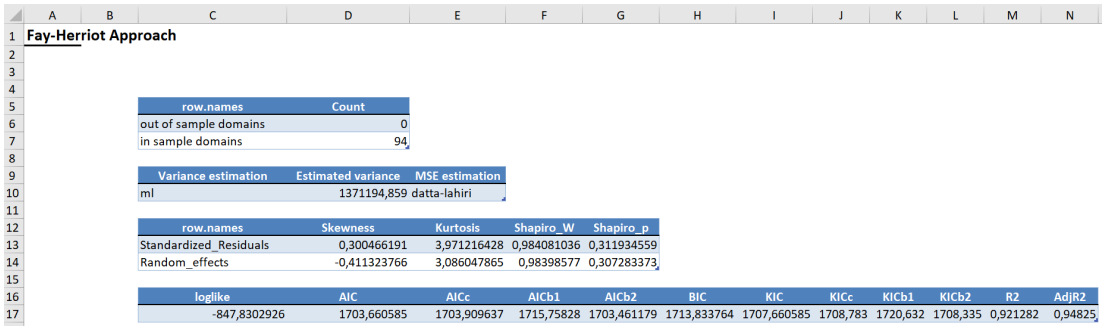
Figure 3.6 provides an insight of the output.

3.4.2 Estimation of the extended area-level models

This section is dedicated to the model building of the extensions of the standard FH model (see Section 3.2.2) implemented in **emdi**. Figure 3.7 in Appendix 3.6 provides an overview of the options that can be chosen and Table 3.6 summarizes which arguments have to be specified for the respective models.

FH model with transformation

If the indicator of interest needs a transformation, either log or arcsin, in addition to the function used in Section 3.4.2, the arguments `transformation` and `backtransformation` must be specified. If, for example, the share of households per area that earn more than the national median income (`MTMED`) is the indicator of interest, an arcsin transformation can be used. The bias-corrected back-transformation `bc` is chosen in the example. Two more arguments are needed when using an arcsin transformation: the name of the variable describing the effective sample sizes (`eff_smpsize`) which needs to be contained in the `combined_data` frame.



Fay-Herriot Approach

row.names	Count
out of sample domains	0
in sample domains	94

Variance estimation	Estimated variance	MSE estimation
ml	1371194,859	datta-lahiri

row.names	Skewness	Kurtosis	Shapiro_W	Shapiro_p
Standardized_Residuals	0,300466191	3,971216428	0,984081036	0,311934559
Random_effects	-0,411323766	3,086047865	0,98398577	0,307283373

loglike	AIC	AICc	AICb1	AICb2	BIC	KIC	KICc	KICb1	KICb2	R2	AdjR2
-847,8302926	1703,660585	1703,909637	1715,75828	1703,461179	1713,833764	1707,660585	1708,783	1720,632	1708,335	0,921282	0,94825

(a)

	A	B	C	D	E	F	G
1	Domain	Direct	Direct_MSE	Direct_CV	FH	FH_MSE	FH_CV
2	Amstetten	14768,56933	926167,3714	0,065163787	14242,04457	599010,649	0,054343165
3	Baden	21995,72487	446534,2852	0,030380095	21616,39582	356586,0515	0,027624784
4	Bludenz	12069,59239	1243265,013	0,092382403	12680,37578	716040,1177	0,066732371
5	Braunau am Inn	10770,53331	1029502,352	0,094205544	11925,8169	643500,244	0,067264547
6	Bregenz	35731,19812	4467316,434	0,059152864	32101,68983	1302156,03	0,035547054
7	Bruck-Mürzzuschlag	23027,3744	1971664,032	0,06097784	22523,49503	906339,1859	0,042267795
8	Bruck an der Leitha	25209,50992	3135150,031	0,070236807	23590,33007	1069157,926	0,043831558
9	Deutschlandsberg	21271,28902	3000062,465	0,081427545	22159,12123	1055862,452	0,046371499
10	Dornbirn	20552,06381	2374522,488	0,074977802	23382,35334	986784,1129	0,042483751

(b)

Figure 3.6: Extract of the Excel spreadsheets created by `write.excel`: (a) `summary` Output, (b) estimation results.

Because of having chosen the bias-corrected back-transformation, the only possible `mse_type` is `boot`, if the MSE estimation is activated.

```
R> fh_arcsin <- fh(
+     fixed = MTMED ~ cash + age_ben + rent + house_allow,
+     vardir = "Var_MTMED", combined_data = combined_data,
+     domains = "Domain", transformation = "arcsin",
+     backtransformation = "bc", eff_smpsize = "n", MSE = TRUE,
+     mse_type = "boot")
```

Spatial FH model

In case the spatial correlation tests conducted in Section 3.4.1 would have indicated a spatial correlation of the domains, a spatial FH model for incorporating the spatial structure in the model could be used. For that the `correlation` has to be set to `spatial` and the proximity matrix exemplarily created in Section 3.4.1 has to be given to the model within the `corMatrix` argument. The possible variance estimation methods are `ml` and `reml`.

```
R> fh_spatial <- fh(fixed = Mean ~ cash + self_empl,
+     vardir = "Var_Mean", combined_data = combined_data,
+     domains = "Domain", correlation = "spatial",
+     corMatrix = eusilcA_prox, MSE = TRUE)
```

Robust FH model

If extreme values could influence the estimation, the application of a robust model might be appropriate. Within the robust framework, package **emdi** allows the user to choose between a standard and a spatial model (defaults to `correlation` = "`no`"). The estimation method must equal `reblup` or `reblupbc` which includes a bias correction that can be modified by the argument `c`. Further, the tuning constant `k` defaults to 1.345 as proposed by Sinha and Rao, 2009 and Warnholz, 2016 and can be changed if desired. The functions of the package **saeRobust** Warnholz, 2018 are utilized for the robust extensions. An exemplary call with pseudolinear MSE estimation looks like this:

```
R> fh_robust <- fh(fixed = Mean ~ cash + self_empl,
+    vardir = "Var_Mean", combined_data = combined_data,
+    domains = "Domain", method = "reblup", MSE = TRUE,
+    mse_type = "pseudo")
```

Measurement error model

If as auxiliary information other data sources than register data, e.g., data from larger surveys or big data sources are used, the ME model should be applicated. For the estimation of the ME model, the model fitting method has to be set to `me` and the only possible MSE estimation method is `jackknife`. The most complex input argument consists of the creation of the MSE array `Ci`. The variability of the auxiliary variables that is taken into account by the ME model is expressed by the variance-covariance matrices per domain (`Ci`). For example, for three covariates a, b and c the array should look like

$$\boldsymbol{C}_i = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathsf{VAR}_i(a) & \mathsf{COV}_i(a,b) & \mathsf{COV}_i(a,c) \\ 0 & \mathsf{COV}_i(a,b) & \mathsf{VAR}_i(b) & \mathsf{COV}_i(b,c) \\ 0 & \mathsf{COV}_i(a,c) & \mathsf{COV}_i(b,c) & \mathsf{VAR}_i(c) \end{pmatrix}, \; i = 1, ..., D.$$

The first row and column contain zeros, because the intercept is considered. The variances and covariances can be computed by standard approaches like for example the Horvitz-Thompson estimator. In R the array is computed by

```
P <- number of covariates
M <- number of areas

Ci_array <- array(data = 0, dim = c(P + 1, P + 1, M))

for(i in 1:M){
  Ci_array[2,2,i] <- Var_a[i]
  Ci_array[3,3,i] <- Var_b[i]
  Ci_array[4,4,i] <- Var_c[i]
```

```
  Ci_array[3,2,i] <- Ci_array[2,3,i] <- Cov_ab[i]
  Ci_array[4,2,i] <- Ci_array[2,4,i] <- Cov_ac[i]
  Ci_array[4,3,i] <- Ci_array[3,4,i] <- Cov_bc[i]
}
```

For the Austrian EUSILC data example, the equalized income can also be explained by a variable of the sample data set. The code below demonstrates how the variance-covariance matrix is created for one covariate (variable `Cash` and its variance `Var_Cash`) and how the final ME model is built.

```
R> P <- 1
R> M <- 94
R>
R> Ci_array <- array(data = 0, dim = c(P + 1, P + 1, M))
R>
R> for(i in 1:M){
+    Ci_array[2,2,i] <- eusilcA_smpAgg$Var_Cash[i]
+  }
R>
R> fh_yl <- fh(fixed = Mean ~ Cash, vardir = "Var_Mean",
+    combined_data = eusilcA_smpAgg, domains = "Domain",
+    method = "me",  Ci = Ci_array, MSE = TRUE,
+    mse_type = "jackknife")
```

3.5 Conclusion and outlook

In this paper, we have presented how the **emdi** package version 1.1.7 has been extended by various area-level models. Besides the well-known FH model, adjusted variance estimation methods and transformation options are offered to the user. In addition, spatial, robust, and ME model extensions of the standard model allow the user to address various issues that arise in practical data applications. All of these methods can be estimated conveniently by using a single function that provides EBLUP and the respective MSE estimates to measure their precision. Especially in Section 3.4 it becomes clear that the package does not only contain the estimation of the different SAE models. Instead, it additionally provides user-friendly tools to enable a whole data analysis procedure: 1. starting with model building and estimation, moving on to 2. model assessment and diagnostics, 3. presentation of the results, and finishing with 4. exporting the results to Excel or OpenDocument Spreadsheet.

For future package versions, it is planned to expand the options in the field of area-level models. In some practical applications the incorporation of random effects is redundant. Therefore, an area-level estimator that considers a preliminary testing for the random effects following Molina, Rao, and Datta (2015) will be included. The **emdi** version 2.0.1 accounts for spatial structures of the random effects. Future developments will also account for out-of-sample

EBLUP and MSE estimation for the spatial model proposed by Saei and Chambers (2005) and for temporal and spatio-temporal extensions (Rao and Yu, 1994; Marhuenda, Molina, et al., 2013). For the existing ME model, a bootstrap MSE estimation option will be added to the package since the Jackknife MSE estimator may produce negative MSE estimates (Marchetti et al., 2015). Furthermore, cross-validation options additional to the model assessment via information criteria and the R^2 will be investigated. Lastly, a stepwise variable selection function and a `compare` method for objects of class `model, ebp` are planned.

Acknowledgments

The work of Kreutzmann and Schmid has been 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.

3.6 Area-level model options and input arguments of function fh

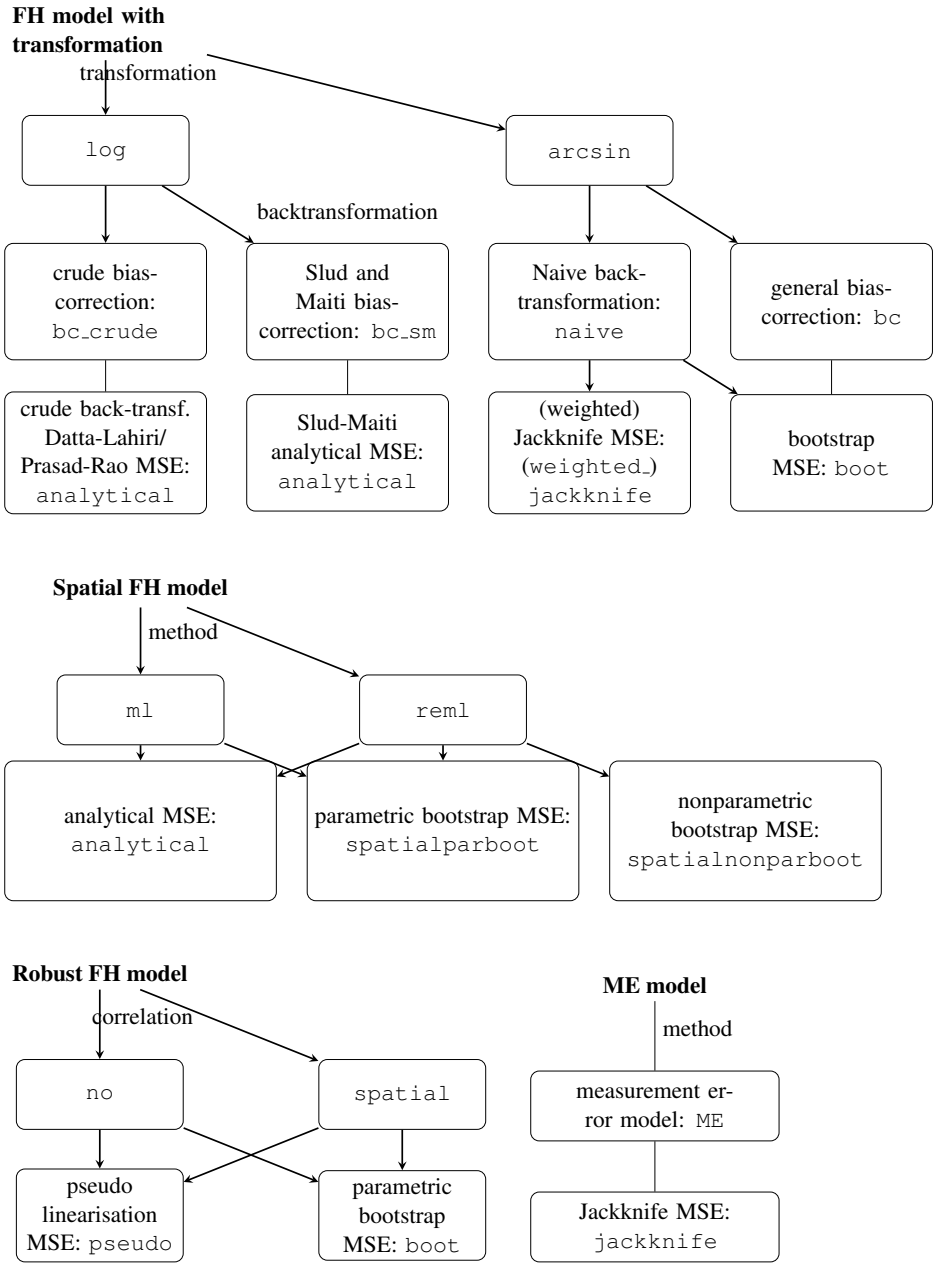


Figure 3.7: Overview of extended area-level models and combinations of estimation methods.

Argument	**FH model**				
	Standard	Transformed	Spatial	Robust	ME
`fixed`	√	√	√	√	√
`vardir`	√	√	√	√	√
`combined_data`	√	√	√	√	√
`domains`	(√)	(√)	(√)	(√)	(√)
`method`	√	√	√	√	√
`interval`	(√)	(√)			
`k`				√	
`c`				√	
`transformation`	√	√	√	√	√
`backtransformation`		√			
`eff_smpsize` (only if `transformation = "arcsin"`)		√			
`correlation`	√	√	√	√	√
`corMatrix` (only if `correlation = "spatial"`)			√	√	
`Ci`					√
`tol`			√	√	√
`maxit`			√	√	√
`MSE`	√	√	√	√	√
`mse_type` (only if `MSE = TRUE`)	√	√	√	√	√
`B`	(√)	√	√	√	
`seed`	(√)	(√)	(√)	(√)	

Table 3.6: Required √ and optional (√) input arguments of function `fh` for the different area-levels models. `B`: Only if bootstrap MSE is chosen. When the standard FH model is applied, `B` is required for the computation of the information criteria by Marhuenda, Morales, et al. (2014) (optionally).

3.7 Output of the model component of an fh object

Name	**Short description**	**Available** for				
		Standard	*Transformed*	*Spatial*	*Robust*	*ME*
`coefficients`	Estimated regression coefficients	√	√	√	√	√
`variance`	Estimated variance of the random effects/ estimated spatial correlation parameter	√	√	√	√	√
`random_effects`	Random effects per domain	√	√	√	√	√
`real_residuals`	Realized residuals per domain	√	√	√	√	√
`std_real_residuals`	Standardized realized residuals per domain	√	√	√	√	√
`gamma`	Shrinkage factors per domain	√	√			√
`model_select`	Model selection and accuracy criteria	√	√	√		
`correlation`	Selected correlation structure of the random effects	√	√	√	√	√
`k`	Tuning constant				√	
`c`	Multiplyer constant for bias correction				√	
`seed`	Seed of the random number generator	√	√	√	√	

Table 3.7: Components of the output component `model` for models of class `fh`.

3.8 Reproducibility

For the computation of the results in this paper we worked with R version 4.0.2 on a 64-bit platform under Microsoft Windows 10 with the installed packages listed in Table 3.8. Using the package **packrat** (Ushey et al., 2018) a snapshot of the corresponding repository was created that is available from the `GitHub` folder (`https://github.com/SoerenPannier/emdi.git`). We suggest the following steps:

- Install Git.
- Create a new project in RStudio.
- Choose checkout from version control and select Git.
- Insert the repository URL: `https://github.com/SoerenPannier/emdi.git`.
- Let **packrat** complete the initialization process.
- Restart RStudio.
- Enter the R command `packrat::restore()`.
- After finishing the installation process all packages are installed as provided in Table 3.8.

Package	Version	Package	Version	Package	Version
BBmisc	1.11	emdi	2.0.1	pkgbuild	1.0.8
BH	1.72.0-3	evaluate	0.14	pkgconfig	2.0.3
DBI	1.1.0	expm	0.999-4	pkgload	1.1.0
HLMdiag	0.3.1	fansi	0.4.1	plyr	1.8.6
KernSmooth	2.23-17	farver	2.0.3	praise	1.0.0
LearnBayes	2.15.1	foreign	0.8-80	prettyunits	1.1.1
MASS	7.3-51.6	formula.tools	1.7.1	processx	3.4.2
Matrix	1.2-18	gdata	2.18.0	ps	1.3.3
MuMIn	1.43.17	ggplot2	3.3.2	purrr	0.3.4
R.cache	0.14.0	glue	1.4.1	raster	3.3-7
R.methodsS3	1.8.0	gmodels	2.18.1	readODS	1.6.7
R.oo	1.23.0	gridExtra	2.3	readr	1.3.1
R.rsp	0.43.2	gtable	0.3.0	rematch	1.0.1
R.utils	2.9.2	gtools	3.8.2	reshape2	1.4.4
R6	2.4.1	highr	0.8	rgeos	0.5-3
RColorBrewer	1.1-2	hms	0.5.3	rlang	0.4.6
RLRsim	3.1-6	isoband	0.2.2	roxygen2	7.1.1
Rcpp	1.0.4.6	knitr	1.29	rprojroot	1.3-2
RcppArmadillo	0.9.900.1.0	labeling	0.3	rstudioapi	0.11
RcppEigen	0.3.3.7.0	laeken	0.5.1	saeRobust	0.2.0
aoos	0.5.0	lattice	0.20-41	scales	1.1.1
assertthat	0.2.1	lifecycle	0.2.0	sf	0.9-4
backports	1.1.8	lme4	1.1-23	simFrame	0.5.3
boot	1.3-25	magrittr	1.5	sp	1.4-2
brew	1.0-6	maptools	1.0-1	spData	0.3.5
callr	3.4.3	markdown	1.1	spdep	1.1-5
cellranger	1.1.0	memoise	1.1.0	statmod	1.4.34
checkmate	2.0.0	mgcv	1.8-31	stringi	1.4.6
class	7.3-17	mime	0.9	stringr	1.4.0
classInt	0.4-3	minqa	1.2.4	testthat	2.3.2
cli	2.0.2	modules	0.8.0	tibble	3.0.1
clipr	0.7.0	moments	0.14	units	0.6-7
coda	0.19-3	munsell	0.5.0	utf8	1.1.4
colorspace	1.4-1	nlme	3.1-148	vctrs	0.3.1
commonmark	1.7	nloptr	1.2.2.1	viridisLite	0.3.0
crayon	1.3.4	openxlsx	4.1.5	withr	2.2.0
deldir	0.1-25	operator.tools	1.6.3	xfun	0.15
desc	1.2.0	packrat	0.5.0	xml2	1.3.2
digest	0.6.25	parallelMap	1.5.0	xtable	1.8-4
e1071	1.7-3	pbapply	1.4-2	yaml	2.2.1
ellipsis	0.3.1	pillar	1.4.4	zip	2.0.4

Table 3.8: Installed packages for the computation of the results in this paper.

Chapter 4

ammlogit: Estimation and Prediction Using an Aggregated Mixed Multinomial Logit Model

4.1 Introduction

Small area estimation (SAE) has become an increasingly important tool to support policy decision making. Therefore, national statistical institutes, such as U.S. Census Bureau, the Office for National Statistics in the UK, and the Italian Statistical Office, have created programs to produce small area estimates. In this context, The term "area" is used in a broader sense than just a geographical region. It may also contain non-geographic dimensions, such as gender or age. However, it is common practice to use the term "area" and the broader term "domain" interchangeably. The rise of SAE in policy making provides additional knowledge about the geographical and/or demographic distribution of relevant indicators. This improved level of information is suitable for adding precision to programs which aim at reducing inequality between areas. Typical indicators include measures of income or wealth, levels of inequality, employment statuses, and rates of poverty. For a comprehensive collection of SAE techniques, see Rao and Molina (2015). Tzavidis, Zhang, et al. (2018) formulate a framework of best practice to obtain small area estimates. The general concept of SAE may be summarized in the principle of "borrowing strength." This means leaning on additional data, in which the actual variable of interest (VOI) is not directly observed, to improve the precision of estimates. For instance, imagine national survey data that includes the VOI, but is too small in size to allow for reliable estimates on a municipal level. Furthermore, there is census data available that is more extensive but does not contain the VOI. In this case, borrowing strength might be achieved by using the survey data for model estimation and combining this model with the census data for prediction. This type of borrowing strength is referred to as "model-based" and is only possible under the assumption that the same model holds for both the survey and census data.

Such model-based SAE methods are generally divided into two major classes, known as the unit level and area level methods. In unit level cases, the data is available for each individual. Hence, both the estimation and the prediction are performed on individual data. The World Bank method (Elbers, Lanjouw, et al., 2003) and the empirical best predictor (Molina and Rao, 2010) are prominent examples in this class. However, for reasons of confidentiality, individual data is not often made available or even published to the research community. It may be possible to instead only access aggregated data. Commonly, aggregation is performed at area level, raising the need for models operating at the same level, most prominently the family of Fay-Herriot models (Fay and Herriot, 1979).

Molina, Saei, et al. (2007) propose a methodology which is a hybrid between these two major classes. The data used is aggregated, but contains subgroups which can be treated as individual observations within each area while they retain their "statistical advantages of aggregated data" (Molina, Saei, et al., 2007). This method will be described in detail in Section 4.2 as it is the foundation of the present paper. Nevertheless, I will give a brief idea of its basic principle beforehand. The aim of Molina, Saei, et al. (2007) is to estimate totals and ratios of three status of labor marked participation. They therefore use data aggregated in subgroups, which divide the data according to domain, sex, and age-group. This results in six distinct observations per area. Three counts are observed in each subgroup, one for each participation. Molina, Saei, et al. (2007) developed a mixed multinomial logit model containing an area level random effect. A closely related approach was undertaken in López-Vizcaíno et al. (2013). In contrast to the previous authors, they propose separate random effects for each modeled status of labor-market participation and used data aggregated on domain-sex-level. The R-package **mme** (Lopez-Vizcaino et al., 2019) implements the latter method in the form of open source software. To the best of my knowledge, the proposed model by Molina, Saei, et al. (2007) is not yet publicly available.

This paper introduces the new R-package **ammlogit**, which implements this method along with R-typical functionality, such as S3-classes and methods. The method described in Molina, Saei, et al. (2007) is extended twofold. Firstly, the predictions may now be based on information on the covariates that differs from the sample, thus allowing for out of sample areas. Secondly, the bootstrap has been adapted to account for the new way of prediction. Additionally, while they were only sampled once and than treated as constant in Molina, Saei, et al. (2007), the bootstrap implemented in **ammlogit** treats the random effects as a random variable and thus resamples them for each bootstrap iteration. The package aims at being user-friendly, trying to close a gap between research and application. In Section 4.2, I will describe the methodology used and also highlight the extension added to the original method. In the subsequent Section 4.3, an exemplary case study on poverty data from Guerrero Mexico will demonstrate the program's functionality.

4.2 Methodology

In this section, I will first generalize the model notation introduced by Molina, Saei, et al. (2007) to the more general case of a VOI Y_j with categories $j = 1, \ldots, J$. The VOI is observed in the domains $d = 1, \ldots, D$ and each domain is divided into I subgroups denoted by $i = 1, \ldots, I$. Consequently, y_{dij} comprises the counts of category j in the subgroup i of area j. Subsequently, $m_{di} = \sum_{j=1}^{J} y_{dij}$ is the sample size of the subgroup i in area j and $n_d = \sum_{i}^{I} m_{di}$ is the sample size of the corresponding domain. Let p_{dij} denote the probabilities of category j within the aggregation di, with $p_{diJ} = 1 - \sum_{j=1}^{J-1} p_{dij}$. The domain-specific random effects are given by u_d with variance φ. With these definitions at hand, Molina, Saei, et al. (2007) assume the vectors $(y_{di1}, \ldots, y_{diJ})$ to be independent across d and i with the multinomial distribution defined by the following probability mass function.

$$f(y_{di1}, \ldots, y_{diJ}|u_d) = \frac{m_{di}!}{\prod_{j=1}^{J} y_{dij}!} \prod_{j=1}^{J} p_{dij}^{y_{dij}} \tag{4.1}$$

Furthermore, a logit link is assumed to connect the covariates and random area effects with the corresponding probabilities:

$$\log(p_{dij}/p_{diJ}) = x_{di}\beta_j + u_d \qquad j = 1, \ldots J-1, \quad u_d \overset{iid}{\sim} N(0, \varphi) \tag{4.2}$$

In this notation, β_j contains the coefficients of the explanatory variables for category j. In this work, the unknown parameters β, u, and φ are estimated by a combination of the penalized quasi-likelihood (PQL) first described by Breslow and Clayton (1993) for β, and u and maximum likelihood (ML) for φ. This so called, PQL-ML method was introduced by Schall (1991) and first used in the context of SAE in Saei and Chambers (2003). The algorithm in the R-package accompanying this manuscript was developed and is described in great detail by Molina, Saei, et al. (2007). I will therefore abstain from repeating the specifics, but refer to Appendix A of the latter work and only promote the general idea in place. In a first step, φ is assumed to be known and the combined log-likelihood of β and u is maximized numerically i.e., using a Newton-Raphson algorithm. Secondly, β and u are considered known and fixed, and a Taylor linearization is used to obtain a maximum likelihood estimate of φ. These two steps are repeated until convergence.

4.2.1 From model fit to prediction

Here, I will distinguish between two situations, both of which are covered in the **ammlogit** package. First, I will give a brief account of the situation described in Molina, Saei, et al. (2007). Second, assuming additional census data is available, I will propose a slightly different approach to making predictions.

Molina describes a situation in which the covariates, as well as the VOI are available from the same sample, additionally only the total population sizes of each subgroup within each

domain M_{di} are known. From this, it is easy to calculate the number of unobserved individuals $m_{di}^{out} = M_{di} - m_{di}$. The objective is now to predict the number of unobserved counts for each VOI class, which are denoted by y_{dij}^{out}.

From the estimation procedure described previously, $\hat{\beta}_j$ and $\hat{u}_d$ are obtained which leads to the following predictor:

$$\hat{y}_{dij}^{out} = m_{di}^{out} \frac{\exp\left(x_{di}\hat{\beta}_j + \hat{u}_d\right)}{1 + \sum_{k=1}^{J-1} \exp\left(x_{di}\hat{\beta}_k + \hat{u}_d\right)} \qquad \text{for } j = 1, \ldots, J-1 \tag{4.3}$$

For the reference category $\hat{y}_{diJ}^{out}$, then follows: $\hat{y}_{diJ}^{out} = m_{di}^{out} - \sum_{j=1}^{J-1} \hat{y}_{dij}^{out}$.

Estimated counts for each subgroup, respectively, each domain, are subsequently given by:

$$\hat{\delta}_{dij} = \hat{y}_{dij}^{out} + \hat{y}_{dij} \quad \text{and} \tag{4.4}$$

$$\hat{\delta}_{dj} = \sum_{i=1}^{I} \hat{\delta}_{dij} = \sum_{i=1}^{I} \hat{y}_{dij}^{out} + \hat{y}_{dij}. \tag{4.5}$$

For this prediction type, Molina, Saei, et al. (2007) suggest a parametric bootstrap for evaluating the mean squared error (MSE) of the estimated δ_{dj}. Their proposed bootstrap treats the random effects as fixed. This assumption is relaxed in the **ammlogit** implementation. To prevent unnecessary duplication, I refer to Molina, Saei, et al. (2007, Section 2.5) for the details of the original bootstrap. However, the approach implemented in **ammlogit** is closely related and will be described in the following subsection 4.2.2.

In this second setting, I assume two data sources: the first one, typically a survey, contains the VOI and the corresponding covariant as also described in the scenario above. Furthermore, additional administrative or census data is available that contains aggregates of the same explanatory variables and grouping as in the survey. This means, not only do we know the population counts for all subgroups M_{di}, but also the population values of x_{id} which I denote as x_{id}^{pop}. In this situation, I suggest using this additional information on the covariates in the prediction by replacing the x_{id} in Equation 4.3 with its population counterpart. This results in

$$\hat{y}_{dij}^{out} = m_{di}^{out} \frac{\exp\left(x_{id}^{pop}\hat{\beta}_j + \hat{u}_d\right)}{1 + \sum_{k=1}^{J-1} \exp\left(x_{id}^{pop}\hat{\beta}_k + \hat{u}_d\right)} \qquad \text{for } j = 1, \ldots, J-1. \tag{4.6}$$

The remaining steps in the prediction, i.e., the calculation of $\hat{y}_{diJ}^{out}$, as well as equation 4.4 and 4.5, remain unchanged. Note that in this setting it is possible to obtain estimates for out-of-sample subgroups and domains. Random effects for out-of-sample areas are set to 0, and for those subgroups $m_{di}^{out} = M_{di}$. This deviation in the point estimation clearly has to be included into the MSE estimation, too. The adapted bootstrap algorithm is described in detail in the

following.

4.2.2 Mean square error estimation

1. Use the original sample data to estimate the model defined in Equations 4.1 and 4.2. Thereby, obtain the parameters $\hat{\beta}_j$ and $\hat{\varphi}$.

2. Generate a vector u^b from a normal distribution with expectation 0 and variance $\hat{\varphi}$. Note: Here, random effects are also generated for out-of sample domains.

3. Generate the bootstrap sample by first calculating the probabilities:

$$\begin{aligned} p_{dij}^b &= \frac{\exp\left(x_{di}\hat{\beta}_j + u_d^b\right)}{1 + \sum_{k=1}^{J-1} \exp\left(x_{di}\hat{\beta}_k + u_d^b\right)} \quad \text{for } j = 1, \ldots, J-1 \text{ and,} \\ p_{diJ}^b &= 1 - \sum_{j=1}^{J-1} p_{dij}^b. \end{aligned}$$

   Then, use the obtained probabilities to generate sample values by means of random numbers generated from the suitable multinomial distribution.

$$y_{di}^b \sim \text{Multinom}\left(m_{di}, p_{dij}^b, \ldots, p_{diJ}^b\right).$$

4. Generate the bootstrap population by first calculating the population probabilities:

$$\begin{aligned} p_{dij}^{b,pop} &= \frac{\exp\left(x_{di}^{pop}\hat{\beta}_j + u_d^b\right)}{1 + \sum_{k=1}^{J-1} \exp\left(x_{di}^{pop}\hat{\beta}_k + u_d^b\right)} \quad \text{for } j = 1, \ldots, J-1 \text{ and} \\ p_{diJ}^{b,pop} &= 1 - \sum_{j=1}^{J-1} p_{dij}^{b,pop}. \end{aligned}$$

   Then, again use the obtained probabilities to generate sample values by means of random numbers obtained from the suitable multinomial distribution.

$$y_{di}^{b,out} \sim \text{Multinom}\left(m_{di}^{out}, p_{dij}^{b,pop}, \ldots, p_{diJ}^{b,pop}\right)$$

   Now, calculate the bootstrap population values $\delta_{dij}^b = y_{dij}^{b,out} + y_{dij}^b$ and $\delta_{dj}^b = \sum_{i=1}^{I} \delta_{dij}^b$.

5. Fit the model from 4.1 and 4.2 on the bootstrap sample data, given by y_{di}^b and x_{di}. Thereby, obtain $\hat{\beta}_j^b$ and $\hat{u}_d^b$ and calculate predicted values by means of equation 4.6, 4.4, 4.5 resulting in $\hat{\delta}_{dij}^b$ and $\hat{\delta}_{dij}^b$.

6. Repeat steps 2 to 5 B-times, using $b = 1, \ldots, B$ as iteration counter.

7. Calculate MSEs by:

$$MSE(\hat{\delta}_{dij}) = B^{-1} \sum_{b=1}^{B} \left(\hat{\delta}^b_{dij} - \delta^b_{dij} \right)^2, \tag{4.7}$$

$$MSE(\hat{\delta}_{dj}) = B^{-1} \sum_{b=1}^{B} \left(\hat{\delta}^b_{dj} - \delta^b_{dj} \right)^2 \tag{4.8}$$

Note: The bootstrap estimation for the case in which no population covariates are available can easily be obtained from the above description by substituting x_{di}^{pop} in step 4 by x_{di} and using Equation 4.3 instead of 4.6 in step 5. Additionally, this bootstrap can be used to calculate MSEs for ratios or other indicators derived from the counts δ_{dij} and δ_{dj}. To do this, an additional indicator needs to be calculated on the estimated and pseudo population values in each iteration. The MSE is then obtained similarly to Equations 4.7 and 4.8. As ratios are frequently at least as relevant as absolute values, the MSE in the package **ammlogit** is automatically calculated for absolute values and ratios.

4.3 Poverty estimation in Guerrero, Mexico, using ammlogit

Even though both Molina, Saei, et al., 2007 and López-Vizcaíno et al., 2013 focus on estimating labor force participation, the possible applications of mulinomial models in SAE are more diverse. Poverty is frequently measured by income or spending. A monetary threshold is defined and, if an individual's income/spending is below this threshold, the agent is classified as poor. A whole family of poverty measures based on this general idea was developed by Foster et al., 1984 (FGT). The FGT-family includes the well known head count ratio (HCR) and poverty gap (PGAP). The underlying poverty line can be defined by absolute or relative terms. Absolute poverty lines are usually defined by a minimal consumption bundle, while relative lines are often expressed as a percentage of the median income within a country or region. However, these measures only include one single dimension of poverty. In a broader sense, poverty might be understood as the shortage or lacking of necessities. The relevant necessity might be monetary, but this might not always be the case as it can also include other factors, e.g. the lack of access to health care or education which are other typical characteristics of poverty. Multidimensional poverty in Mexico is defined by the "National Council for the Evaluation of Social Development Policy" (CONEVAL) (CONEVAL, 2010) as the aggregation of two dimensions: the monetary dimensions measured by income and the deprivation of social rights. The social rights considered in CONEVAL (2010, p.4) are:

1. "educational lag",
2. lack of access to health services,
3. lack of access to social security,
4. housing with inadequate quality or insufficient space,
5. lack of basic housing services, and
6. lack of access to food.

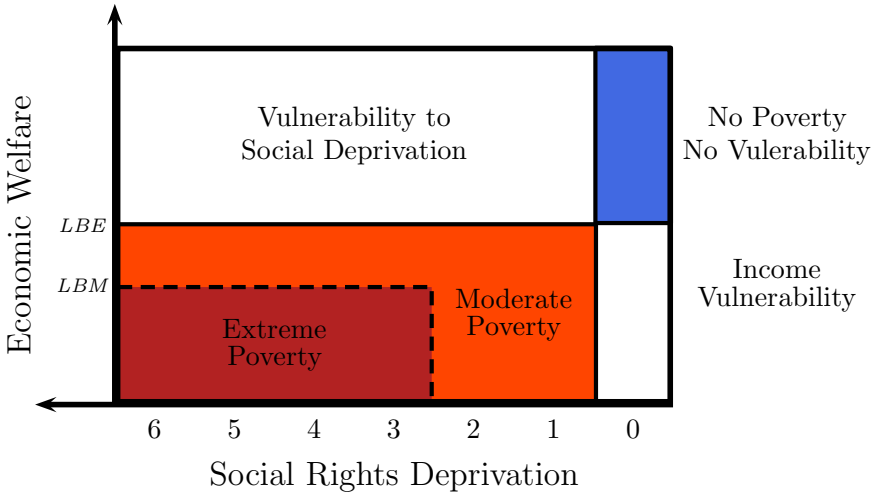


Figure 4.1: CONEVAL definition of multidimensional poverty in Mexico: Individuals are classified into one of five poverty classes. LBE stands for well-being and LBM denotes the minimum threshold of well-being. Graphic reproduced from CONEVAL (2010, p. 5) with minor changes.

This then results in Figure 4.1 which illustrates the five resulting poverty classes. Figure 4.1 illustrates the necessity for an unordered multinomial model for poverty classification. The three antidiagonal classes "No Poverty/No Vulnerability" (NN), "Extreme Poverty" (EP), and "Moderate Poverty" (MP) possess an intrinsic order that would allow for using more parsimonious ordinal methods, such as cumulative logit models. However, the two remaining classes, "Vulnerability to Social Deprivation" (SV) and "Income Vulnerability" (IV) cannot be ordered as it is not inherently clear which of the two is worse to the individual or society as a whole, an unordered modeling approach is hence required.

The R-Package **ammlogit** is centered around the similarly named function `ammlogit`. This function is the package's workhorse as it estimates the model as described in Equations 4.1 and 4.2 and also produces point estimates for domains and subgroups; last but not least, it performs the bootstrap procedures described in Section 4.2.2.

`ammlogit` works in two modes, the model estimation mode and the prediction mode. They do not need to be explicitly selected, but are internally derived from the input arguments. If the arguments `totals` and `population` are left blank or set to `NULL`, only a model estimation is performed. If no `population` is provided, but `totals` are, the prediction will be performed as described in Molina, Saei, et al. (2007). Last but not least, if a `population` is passed to the function, the prediction will be performed as suggested in this paper (cf. Equation 4.6)

4.3.1 Data preparation

The function `ammlogit` expects all data to be provided in the "tidy" (Wickham et al., 2014) respectively "long" format. In this context, "tidy" means each row to contain a single ob-

servation, and each column holds a single variable. However, the data may be provided in a standard `data.frame` and does not need to be a **tidyverse** (Wickham et al., 2019) `tibble`. An illustration of the data structure in this specific example is shown below.

```
R>  head(x = aggrNumericSurvLong[c("idD", "subGroup",
+     "PovertyClasses", "PovertyTotals", "tam_hog", "edad")],
+     12)
```

```
# A tibble: 12 x 6
idD      subGroup PovertyClasses PovertyTotals tam_hog  edad
<chr>    <chr>    <fct>                  <dbl>   <dbl> <dbl>
 1 12001 0_1      EP                        76    5.46  9.93
 2 12001 0_1      MP                       149    5.46  9.93
 3 12001 0_1      SV                        75    5.46  9.93
 4 12001 0_1      IV                        11    5.46  9.93
 5 12001 0_1      NN                        36    5.46  9.93
 6 12001 0_2      EP                        81    4.72 37.0
 7 12001 0_2      MP                       193    4.72 37.0
 8 12001 0_2      SV                       173    4.72 37.0
 9 12001 0_2      IV                        12    4.72 37.0
10 12001 0_2      NN                        92    4.72 37.0
11 12001 0_3      EP                        10    4.16 73.7
12 12001 0_3      MP                        13    4.16 73.7
```

Note: Each subgroup is repeated five times, once for each poverty class, as each count is considered a single observation. Looking at the same output in the census data set:

```
R>  head(x = aggrNumericCens[c("idD", "subGroup",  "tam_hog",
+     "edad", "Totals")], 9)
```

```
# A tibble: 9 x 5
idD      subGroup tam_hog  edad Totals
<chr>    <chr>      <dbl> <dbl>  <int>
 1 12001 0_1         5.27  14.5  40043
 2 12001 0_2         4.72  36.5 183020
 3 12001 0_3         4.50  73.2  11974
 4 12001 1_1         5.34  14.7  38617
 5 12001 1_2         4.66  36.7 196481
 6 12001 1_3         4.21  74.1  13986
 7 12002 0_1         7.21  14.4   1205
 8 12002 0_2         6.40  37.3   2870
 9 12002 0_3         4.55  74.5    328
```

Note that the variables “PovertyClasses” and “PovertyTotals” are missing as they are not ob-

served in the census. As a result, there only exists a single line for each subgroup in the census. Compared to the survey data, the census `data.frame` contains one additional column, `Totals`. This column contains the best available estimation for the population size in each subgroup. Even though the data here appears to be already sorted by domain (idD) and subgroup, this is no necessity, as both the survey data as well as the census data will be sorted internally by the `ammlogit` function.

4.3.2 Estimation

The model used in this example was obtained by fitting non-mixed models for each single category and using R-`stats`' `step`-function with BIC optimization to find a parsimonious model. The combined model then consists of the union of all selected variables.

```
R>  fixedForm <- PovertyTotals ~ sexoMujer + rururbRural +
+     veintil_ing + muj_hog + ic_sbvCon_carencia +
+     remesasRecibe + pcpering + clase_hogCompuesto +
+     inglabpc_esc + tam_hog + edad + jubiRecibe +
+     dhind_parentSí_tiene + años_escolaridad
```

Table 4.1: Arguments of `ammlogit`

Argument	Mandatory	Default	Function
fixed	✓	✗	Formula of the fixed effects
domain	✓	✗	Name of the column defining the domains
grouping	✓	✗	Name of the column defining the subgroups
lvlName	✓	✗	Name of the column containing the level/category names
sample	✓	✗	The sample data
totals		✗	Name of the totals column in the population data or a separate data.frame containing the totals column and the domain and subgroup columns
population		✗	The census data
BS	✓	TRUE	A logical value indicating if the rmse of the predictions should be calculated
B		50	If BS is true, B defines the number of bootstraps used
seed		123	The seed provided to the random number generation of the bootstrap
opt.control		NULL	A named list providing some control over the optimization procedure

The model estimation is performed with the command `ammlogit`. As can be seen in the

function call below, **ammlogit** accepts numerous arguments which are summarized in Table 4.1.

```
R>  ammModel <- ammlogit(fixed = fixedForm, domain = "idD",
+     grouping = "subGroup", sample = aggrNumericSurvLong,
+     totals = "Totals", population = aggrNumericCens,
+     lvlName = "PovertyClasses" ,BS = TRUE, B = 200,
+     seed = 123)
```

As `ammModel` is an `amml` `S3`-object a couple of typical methods are implemented and provided by the package **ammlogit**.

First, a `print` method shows a very short summary, including the call, the levels of the VOI, and the number of domains used in the model estimation.

```
R>  print(ammModel)
```

```
Call:
ammlogit(fixed = PovertyTotals ~ sexoMujer + rururbRural +
  veintil_ing + muj_hog + ic_sbvCon_carencia + remesasRecibe +
  pcpering + clase_hogCompuesto + inglabpc_esc + tam_hog +
  edad + jubiRecibe + dhind_parentSí_tiene + años_escolaridad,
  domain = "idD", grouping = "subGroup",
  lvlName = "PovertyClasses", sample = aggrNumericSurvLong,
  totals = "Totals", population = aggrNumericCens, BS = TRUE,
  B = 200, seed = 123)

Categories:
"MP" "NN" "EP" "IV" "SV"

Number of domains used:
40
```

Additional information can be extracted by the `summary` method. The output of the summary is grouped into three components. First, the call is printed, then the model summary is given, and finally a summary of the prediction is shown.

The object created by the above call is an `S3` object of class `amml`. It contains the function call and three further elements:

- `model`: a named list containing model estimates and design matrices.
- `baseData`: a named list containing general information, e.g., the totals, levels, in-sample domains, all domains.
- `prediction`: a named list containing the point estimates (`PointEstimates`) and

the bootstrap MSE (`MSE`) estimates.

The model summary shows the reference category and a coefficient table containing the estimated coefficients, their estimated standard-errors, the test statistic, and the corresponding p-value. As a separate coefficient needs to be estimated for all categories (except the reference), this table can in practice become very large.

In the given example, moderate poverty is treated as the reference category. By interpreting one of the variables in the different categories, I have analyzed the impact of an area being rural in contrast to being urban. This is measured in the variable `rururbRural`, which is 1 in rural areas and 0 in urban areas. This variable is significant for most of the categories and it can be observed that living in an urban area ($1 - rururbRural$) increases the probability of being neither poor nor vulnerable relative to the probability of being moderately poor. On the other hand, if an area is rural and all other things are held constant, the probability of extreme poverty decreases relative to the probability of moderate poverty. Income vulnerability is the only category for which `rururbRural` has no significant impact and this seems plausible. Furthermore, coderururbRural increases the probability for vulnerability to social deprivation relative to the probability of moderate poverty, which, as infrastructure is generally better in urban areas compared to rural ones, is also intuitive.

The last row of the model-summary informs the user of the estimated random effect variance, which is approximately $\hat{\varphi} = 0.19$

```
R> summary(firstEst)

Call:    ammlogit(fixed = PovertyTotals ~ sexoMujer + rururbRural +
   veintil_ing + muj_hog + ic_sbvCon_carencia + remesasRecibe +
   pcpering + clase_hogCompuesto + inglabpc_esc + tam_hog + edad +
   jubiRecibe + dhind_parentSí_tiene + años_escolaridad,
   domain = "idD", grouping = "subGroup", lvlName = "PovertyClasses",
   sample = aggrNumericSurvLong, totals = "Totals",
   population = aggrNumericCens, BS = TRUE, B = 200, seed = 123)

############################################################

Model summary:
Reference category:         MP

Estimated coefficients:
                          Coefficient  Std.D.     Test    P-Value
NN: (Intercept)          -1.3678e+00  1.2788e+00 -1.0696 0.2848149
NN: sexoMujer            -6.0962e-02  1.7485e-01 -0.3487 0.7273468
NN: rururbRural          -1.7404e+00  4.5311e-01 -3.8409 0.0001226 ***
NN: veintil_ing           6.1439e-02  8.1320e-02  0.7555 0.4499381
NN: muj_hog               8.6899e-02  2.4279e-01  0.3579 0.7204060
NN: ic_sbvCon_carencia    1.7773e-01  5.1382e-01  0.3459 0.7294187
```

```
NN: remesasRecibe        -9.2816e-01  7.6054e-01 -1.2204 0.2223150
NN: pcpering             -6.9331e-03  1.3566e-02 -0.5111 0.6093060
NN: clase_hogCompuesto   -2.4399e-01  2.7340e+00 -0.0892 0.9288887
NN: inglabpc_esc          2.3487e-04  1.4312e-04  1.6410 0.1007910
NN: tam_hog              -2.4045e-01  1.6916e-01 -1.4215 0.1551829
NN: edad                 -2.3271e-05  6.4473e-03 -0.0036 0.9971202
NN: jubiRecibe            2.7179e+00  9.9669e-01  2.7269 0.0063925 **
NN: dhind_parentSí_tiene  1.0594e+00  9.6468e-01  1.0982 0.2721352
NN: años_escolaridad      6.1791e-02  4.7367e-02  1.3045 0.1920585
EP: (Intercept)           2.4260e-01  7.3894e-01  0.3283 0.7426748
EP: sexoMujer             3.9080e-02  1.0413e-01  0.3753 0.7074431
EP: rururbRural          -7.7676e-01  2.8296e-01 -2.7452 0.0060483 **
EP: veintil_ing          -2.7819e-01  5.3074e-02 -5.2416 1.592e-07 ***
EP: muj_hog               1.3927e-02  1.3706e-01  0.1016 0.9190659
EP: ic_sbvCon_carencia    1.6891e+00  3.5461e-01  4.7632 1.905e-06 ***
EP: remesasRecibe        -1.9983e+00  4.4122e-01 -4.5289 5.928e-06 ***
EP: pcpering             -1.3560e-02  6.4954e-03 -2.0876 0.0368339 *
EP: clase_hogCompuesto   -3.8160e+00  1.7357e+00 -2.1985 0.0279115 *
EP: inglabpc_esc          4.2230e-04  1.4621e-04  2.8883 0.0038730 **
EP: tam_hog               3.3069e-01  1.0084e-01  3.2794 0.0010402 **
EP: edad                  1.1205e-02  3.7785e-03  2.9655 0.0030220 **
EP: jubiRecibe           -1.1162e+00  8.6410e-01 -1.2918 0.1964368
EP: dhind_parentSí_tiene -1.3151e-01  6.3566e-01 -0.2069 0.8360910
EP: años_escolaridad     -2.6182e-02  2.5590e-02 -1.0232 0.3062313
IV: (Intercept)          -3.8583e+00  2.2053e+00 -1.7495 0.0802032 .
IV: sexoMujer            -1.4809e-01  2.9981e-01 -0.4940 0.6213374
IV: rururbRural          -1.1057e+00  8.1235e-01 -1.3611 0.1734836
IV: veintil_ing           4.5492e-01  1.2636e-01  3.6003 0.0003179 ***
IV: muj_hog               4.9010e-01  3.8378e-01  1.2770 0.2015892
IV: ic_sbvCon_carencia    2.6543e-01  7.2981e-01  0.3637 0.7160858
IV: remesasRecibe        -1.2415e+00  1.2793e+00 -0.9704 0.3318393
IV: pcpering              1.5170e-02  2.1567e-02  0.7034 0.4818176
IV: clase_hogCompuesto   -1.6945e+00  6.7111e+00 -0.2525 0.8006652
IV: inglabpc_esc         -6.4117e-04  3.1153e-04 -2.0581 0.0395783 *
IV: tam_hog              -7.6295e-01  2.7321e-01 -2.7926 0.0052293 **
IV: edad                 -5.1215e-03  9.9870e-03 -0.5128 0.6080806
IV: jubiRecibe           -2.5999e-01  1.5530e+00 -0.1674 0.8670458
IV: dhind_parentSí_tiene -3.1463e-01  1.2480e+00 -0.2521 0.8009512
IV: años_escolaridad     -7.8992e-02  6.4240e-02 -1.2296 0.2188331
SV: (Intercept)          -3.6067e+00  8.8515e-01 -4.0746 4.608e-05 ***
SV: sexoMujer             3.3447e-03  1.2324e-01  0.0271 0.9783483
SV: rururbRural           8.7830e-01  3.1318e-01  2.8044 0.0050406 **
SV: veintil_ing           3.1393e-01  5.1352e-02  6.1134 9.755e-10 ***
SV: muj_hog               2.0364e-02  1.6128e-01  0.1263 0.8995228
SV: ic_sbvCon_carencia    1.3159e+00  3.7249e-01  3.5326 0.0004115 ***
SV: remesasRecibe         2.3569e-01  4.6417e-01  0.5078 0.6116150
```

```
SV: pcpering               9.4022e-03  7.9815e-03  1.1780 0.2387976
SV: clase_hogCompuesto     5.5124e-02  2.0345e+00  0.0271 0.9783842
SV: inglabpc_esc           1.5984e-04  1.2295e-04  1.3000 0.1935863
SV: tam_hog               -2.8374e-01  1.1531e-01 -2.4607 0.0138656 *
SV: edad                   3.8542e-03  4.1196e-03  0.9356 0.3495026
SV: jubiRecibe            -8.0160e-01  7.8671e-01 -1.0189 0.3082387
SV: dhind_parentSí_tiene -9.8007e-01  6.4605e-01 -1.5170 0.1292638
SV: años_escolaridad      -6.0872e-02  2.6809e-02 -2.2706 0.0231706 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated random effect variance:         0.1866424

############################################################

Prediction summary:
Number of in-sample domains:         40
Number of out-of-sample domains:     41

Summary of estimated domain totals:
                  MP            NN          EP            IV            SV
Min.       370.2572     7.092501   567.3847 3.072422e-01     33.00998
1st Qu.   1601.6865    65.258244  2079.6459 1.147422e+01    505.71290
Median    3133.7437   224.673793  4175.0423 6.659589e+01   1882.63673
Mean      7229.8207  2005.058404  6086.1649 7.219744e+02   7887.64823
3rd Qu.   6631.8003   641.223678  7843.9889 2.564054e+02   5301.33115
Max.    119414.8053 75160.231397 31082.5834 2.824031e+04 230223.06622

Summary of the corresponding rel. RMSE:
            MP     NN     EP      IV     SV
Min.     7.501 14.313  2.961  24.651  5.984
1st Qu. 19.494 24.030  8.498  37.857 13.469
Median  26.270 27.491 11.631  45.749 17.660
Mean    26.284 30.569 12.185  57.959 16.814
3rd Qu. 32.072 35.071 15.560  63.840 19.454
Max.    62.666 75.253 22.727 256.258 36.693
```

The prediction summary consists again of two parts, a summary of count estimates on domain level in each category and their corresponding estimated relative root mean squared errors in percentages (rRMSE_{dj}), where

$$\mathrm{rRMSE}_{dj} = \frac{100\,\sqrt{\mathrm{MSE}(\hat{\delta}_{dj})}}{\hat{\delta}_{dj}}.$$

In this example, the range of point estimates are spread on a wide range, stretching from ap-

proximately 0 (min. IV) to more than 200,000 (max. SV). The rRMSE also differs strongly, inside as well as between categories. Extreme poverty seems to be estimated with satisfying precision ranging only between 3 and 23%. On the other hand the rRMSE of IV indicates that the corresponding point estimates should not be trusted, considering the extreme maximum rRMSE of 256%. This very high value might be explained by either a large MSE or a very small point estimate. This particular value will be investigated in the next step.

For this purpose, the domain estimates need to be extracted. This is easily achieved with **ammlogit**'s generic estimators function. `estimators` is written as a S3-method to avoid masking problems with other packages using the same function name to perform a similar purpose e.g., the R-package **emdi** Kreutzmann, Pannier, et al., 2019.

```
R>  ammEsts <- estimators(ammModel, domain = TRUE,
+     ratio = FALSE, category = "IV", RMSE = TRUE,
+     rRMSE = TRUE)
```

`estimators` accepts the arguments domain, ratio, category, RMSE, and rRMSE. Domain switches between estimates on domain and subgroup level, ratio decides whether estimates of totals or ratios should be returned, and category selects the category of the VOI for which results are extracted. This last argument can also be set to "all", returning all categories. The last two arguments select the estimates of precision that should be returned.

```
R>  class(ammEsts)
```

```
"ammlEsts"   "data.frame"
```

`ammEsts` has a class of its own (`ammlEsts`), but also inherits the `data.frame` class. This means that, even though there are specially tailored methods for the `ammlEsts` class, additionally standard `data.frame`-methods, such as `head`, `tail`, etc., are natively available.

```
R>  head(ammEsts)
```

```
                IV     IV_RMSE IV_rRMSE
12001 28240.313740 7450.794353 26.38354
12002    10.427550    7.865947 75.43428
12003   110.711407   49.822120 45.00179
12004     3.431882    1.657868 48.30785
12005    68.295761   27.433862 40.16920
12006    51.551097   19.545280 37.91438
```

Now that the IV estimates are extracted, the row containing the maximum of `IV_rRMSE` can be printed by the following command.

```
R>  ammEsts[which.max(ammEsts[["IV_rRMSE"]]), ]

              IV   IV_RMSE IV_rRMSE
12078 0.3072422 0.7873322 256.2578
```

In this particular case, the error in area 12001 seems to be much more grave, with an RMSE of over 7000, which only leads to a relative RMSE of 26% in contrast, the extreme rRMSE of 256% in area 12078 only results from an RMSE of 0.79.

4.3.3 Visualization

After obtaining the estimates, one next logical step is to visualize these in form of a map. **ammlogit** assists the user in this, provided a spatial polygon of class `sf` (Pebesma, 2018) is available. Generally, domain identifiers in spatial polygons differ from those existing in survey or census data and deriving a universally applicable matching strategy is extremely challenging (Wickham, 2016). The **ammlogit** solution to this is to allow for a so called mapping table, containing the census domain IDs in one column and the corresponding spatial frame IDs in another.

```
R>  head(spatialPolygone[1:6])

Simple feature collection with 6 features and 10 fields
geometry type:  MULTIPOLYGON
dimension:      XY
bbox:           xmin: -100.8357 ymin: 16.68403
                xmax: -98.19347 ymax: 18.23988
geographic CRS: WGS 84
# A tibble: 6 x 11
   ID_0 ISO   NAME_0  ID_1 NAME_1  ID_2 NAME_2 VARNAME_2
  <int> <chr> <chr>  <int> <chr>  <int> <chr>  <chr>
1   143 MEX   Mexico  1815 Guerr~ 20492 Acapu~ Acapulco~
2   143 MEX   Mexico  1815 Guerr~ 20493 Ahuac~ NA
3   143 MEX   Mexico  1815 Guerr~ 20494 Ajuch~ Ajuchitl~
4   143 MEX   Mexico  1815 Guerr~ 20495 Alcoz~ NA
5   143 MEX   Mexico  1815 Guerr~ 20496 Alpoy~ NA
6   143 MEX   Mexico  1815 Guerr~ 20497 Apaxt~ NA
# ... with 1 more variable: geometry <MULTIPOLYGON [°]>
```

This is also the case in the example provided here, as the variable `ID_2` does not match the `idD` domain identifier existing in the data. Such tables may be created completely manually in external spreadsheets, but it is often worth first investing some time to research if a suitable table (at least partially) may already be available via a third data source.

The mapping table's structure for the Guerrero case can be seen below. The NAME_2 column corresponds with the polygon data, and the idD column with the census domain identifier.

```
R>  head(mappingTable)
```

```
# A tibble: 6 x 2
  NAME_2                  idD
  <chr>                   <chr>
1 Acapulco de Juárez      12001
2 Ahuacuotzingo           12002
3 Ajuchitlán del Progreso 12003
4 Alcozauca de Guerrero   12004
5 Alpoyeca                12005
6 Apaxtla                 12006
```

EP is estimated with reasonable accuracy and is also the most urgent kind to eradicate. Therefore, in the next step, the ratios of extreme poverty and their accuracy are analyzed and visualized.

```
R>  EPratio <- estimators(ammModel, domain = TRUE, ratio = TRUE,
+     category = "EP", RMSE = TRUE, rRMSE = TRUE)
R>  summary(EPratio)
```

```
       EP              EP_RMSE            EP_rRMSE
Min.   :0.05757   Min.   :0.006361   Min.   : 2.961
1st Qu.:0.17981   1st Qu.:0.026436   1st Qu.: 8.498
Median :0.40457   Median :0.040030   Median :11.631
Mean   :0.43351   Mean   :0.043132   Mean   :12.185
3rd Qu.:0.67834   3rd Qu.:0.058597   3rd Qu.:15.560
Max.   :0.94761   Max.   :0.081927   Max.   :22.727
```

It is clear that the prevalence of extreme poverty varies strongly between Guerrero's municipalities, as it ranges from 95% to 5% and thus nearly covers the complete possible range. In order to gain further understanding of the distribution of EP in Guerrero, the mapPlot-function can be used with relative ease.

```
R>  mapPlots <- mapPlot(object = EPratio,
+     spatialFrame = spatialPolygone,
+     mappingTable = mappingTable, idInMap = "NAME_2")
```

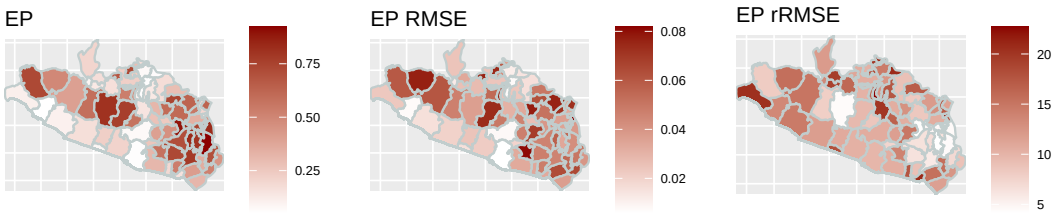


Figure 4.2: Ratio of extreme poverty in Guerrero, Mexico, corresponding RMSE and relative RMSE in percentages.

The above call produces the maps shown in Figure 4.2 and returns a list containing the `ggplot`-type objects (Wickham, 2016) for each graphic. Note that, instead of the above call, the following call would have resulted in the exact same plots, as `mapPlot` is a generic and implemented for the two classes, `ammlEsts` and `ammlogit`:

```
R>  mapPlots <- mapPlot(object = ammModel,
+     spatialFrame = spatialPolygone,
+     mappingTable = mappingTable, idInMap = "NAME_2",
+     domain = TRUE, ratio = TRUE, category = c("EP"),
+     RMSE = TRUE, rRMSE = TRUE)
```

Extreme poverty is strongly clustered, with the strongest cluster located in south-east and a second, line-like cluster located in the north-west. When comparing the pattern of poverty with the geographical information about Guerrero, it can be seen that these two clusters correspond with the sparsely populated mountainous parts of Guerrero. Furthermore, Chilpancingo de los Bravo (Guerrero's capital city) and the main highway which connects Acapulco de Juárez (largest city in Guerrero) with Mexico City (Mexico's capital city) correspond to the area separating these two clusters. The coastal municipalities, which create the south-western boundary of Guerrero, show comparatively small percentages of extreme poverty. This might indicate a relative wealth of these regions, stemming from tourist hot spots, such as Acapulco, Ixtapa and Zihuatanejo. These findings are in line with Rojas-Perilla et al. (2020) who find similar patterns when investigating income based poverty measures, such as head count ratio and poverty gap (Foster et al., 1984).

The `mapPlot` function allows for some customization of the resulting plots. Nonetheless, the most convenient way to achieve another style of presentation is to use the layer-type functionality of `ggplot`-objects and the returned list of plots. Lets say, the title should explicitly state "Extreme poverty", the legend named "Ratio" should range exactly from zero to one, the gray background grid should be removed, and a black and white version is needed to print it. This can all be achieved by overlaying the original plot:

```
R>  mapPlots[["EP"]] + ggtitle("Extreme poverty") +
+     theme_void() + scale_fill_gradient(name = "Ratio",
+     low = "white", high = "black", limits = c(0,1))
```

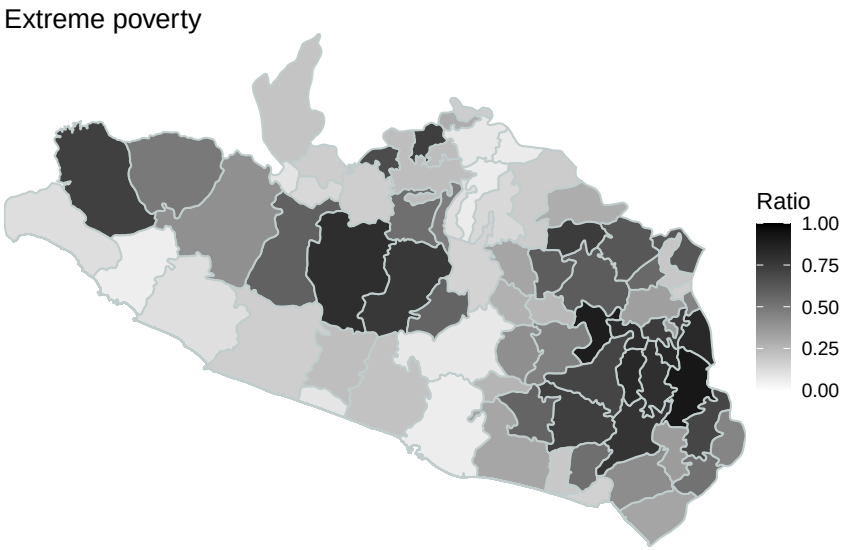


Figure 4.3: Customized plot, showing ratios of extreme poverty in Guerrero.

This allows the user to create customized publication-quality spatial visualizations of the estimates obtained through **ammlogit** with only basic knowledge in R and **ggplot2**. Especially, the hurdle of handling spatial objects is removed. However, if the agent is sufficiently proficient in handling spatial objects himself, changing the `returnValue` argument in `mapPlot` alters the function's behavior to returning the spatial frame, enriched with the data of the chosen estimates. This allows users to build their own plots from scratch.

4.4 Conclusion and future developments

In this paper, I have demonstrated how the **ammlogit** package implements the estimation of a mixed multinomial logit model on aggregated data and facilitates small area predictions based on this model. As a result, applying the method proposed by Molina, Saei, et al. (2007) is made much less complex and, as such, this work may help in making it much easier to use for researchers in more applied fields as well as practitioners at national statistical institutes. **ammlogit** thus narrows the gap between model-developing and model-applying research. It also assists the user in terms of model assessment, result visualization, and result exportation.

In the future, additional methods and features will be added to this package. The model proposed by López-Vizcaíno et al. (2013) i.e., a similar method but with separate random effects per category of the response variable, is a further natural candidate to expand the package's applicability.

Part III

Asymptotics

Chapter 5

Asymptotic Distribution of Regression Quantiles in a Mixed Effects Model

5.1 Introduction

Quantile regression is widespread within many fields of research. Its applications range from economics to climate research, and the method is frequently applied in medicine and biology. In its column "Points of Significance", Nature Methods even dedicated a prominent article (Das et al., 2019) on it. In contrast to ordinary least squares regression, quantile regression relies on the proportion of observations being below a certain value instead of distance. This property makes it robust to outliers, and, in the case of the median, enables a more stable estimation of the center of the data. Furthermore, apart from the error terms to be conditionally independent and identically distributed (iid), quantile regression does not require any distributional assumptions on the response. Regressions for quantiles other than the median allow for an assessment of the spread, and visualization, e.g., through boxplots.

Quantile regression was originally proposed by Koenker and Bassett Jr (1978). Since then, several extensions and properties have been derived. The equivalence to an asymmetric Laplacian linear model has been of, and is in this work, of particular relevance. This property was first shown for a median model in Jung (1996), and later extended for all quantiles in Koenker and Machado (1999). This representation of the quantile regression by an asymetric Laplacian linear model allows for an estimation via maximum likelihood type methods. There is extensive theory surrounding maximum likelihood estimators and their properties, hence the estimator can be shown to be asymptotically normally distributed and therefore consistent.

In many applications, the iid assumption on observation is not justified. This is a general problem in repeated measurements, longitudinal, or clustered data. In order to address this frequent issue, the linear mixed model was introduced by Geraci and Bottai (2007) and Geraci and Bottai (2014). Both use the equivalence of the quantile regression to an asymetric Laplacian linear regression and suggest an estimation using maximum likelihood. Unfortunately,

there is no closed form solution maximizing the log-likelihood and, henceforth, numerical optimization algorithms need to be applied. Geraci and Bottai (2007) use an EM-algorighm (McLachlan and Krishnan, 2007) whereas Geraci and Bottai (2014) suggest a two-stage maximum likelihood approach which includes Gaussian-quadrature (Pinheiro and Chao, 2006). The R-package lqmm (Geraci, 2014) enables practitioners to use mixed quantile regression models as described in Geraci and Bottai (2014), with a low hurdle.

A first approach on proving the consistency of the maximum likelihood estimator was conducted in Weidenhammer (2016). Grounded on the proof of consistency for linear mixed models (Miller, 1977; Pinheiro, 1994), Weidenhammer consequently applies the Weiss' Theorem Weiss (1971) and Weiss (1973) in her proof.

The remaining parts of the paper are organized as follows. In Section 5.2, we will introduce the notation of the linear quantile mixed model used in this paper, show the assumptions on the design that we have used, and formally introduce our main result: the asymptotic distribution of the estimated parameters. Subsequently, in Section 5.3, we will calculate suitable derivatives of our $\log$-likelihood. Based on these computations, in Section 5.4, we will then carry out the proof of our main theoretical result, Theorem 5.2.1. As the proof itself is rather heavy from the computational point of view, we will also make some remarks about the proof's structure at the beginning of that section to highlight the conceptual ideas. The final section, 5.5, is devoted to a small-scale simulation study which complements the theory.

5.2 Notation and assumptions

5.2.1 The linear quantile mixed model

We give a more detailed description of the model to be considered throughout this paper. The model itself is due to Geraci and Bottai (2007) and Geraci and Bottai (2014).

Let $M = M(N)$ denote the number of clusters, $n_i = n_i(N)$ the size of the ith cluster and $N = \sum_{i=1}^{M} n_i$ the overall size of the sample. Fix $\tau \in (0,1)$. Furthermore let $(\Omega, \mathcal{F}, \mathsf{P})$ be some probability space. The response and the error term are $\mathbf{R}^N$-valued random variables, which shall be denoted by Y and ε respectively. Let p (resp. q) be the number of fixed (resp. random) effects to be incorporated in our model. The (non-stochastic) design matrix $X \in \mathbf{R}^{N\times p}$ connecting the fixed effects $\beta_\tau \in \mathbf{R}^p$ with the response is composed of the (non-stochastic) design matrices $X_i \in \mathbf{R}^{n_i \times p}$ on cluster level, i.e., $X = (X_1, \dots, X_M)$.

The random effects are given by some $\mathbf{R}^{Mq}$-valued random variable U, the corresponding (non-stochastic) design matrix connecting U with Y by some block-diagonal matrix $Z = \operatorname{diag}(Z_1, \dots, Z_M) \in \mathbf{R}^{N\times Mq}$, with $Z_i \in \mathbf{R}^{n_i \times q}$ for all $i = 1, \dots, M$. Furthermore we decompose $Y = (Y_1, \dots, Y_M)$, $(\varepsilon_1, \dots, \varepsilon_M)$ and also $U = (U_1, \dots, U_M)$, each one in the obvious sense. With these conventions we now impose the linear mixed model

$$Y = X\beta_\tau + ZU + \varepsilon, \tag{5.1}$$

i.e., on cluster level

$$Y_i = X_i\beta_\tau + Z_iU_i + \varepsilon_i, \quad i = 1, \dots, M. \tag{5.2}$$

In the following, we assume $(U_i)_{i=1,\dots,M}$ to be an iid family, each component being distributed according to some absolutely continuous cdf such that U and ϵ are independent. Additionally, the density of U_i with respect to Lebesgue measure shall be completely determined by some symmetric positive definite (spd) matrix $\Psi_\tau \in \mathbf{R}^{q\times q}$ representing the covariance matrix of each U_i. Writing now $\varepsilon_i = (\varepsilon_{ij})_{j=1,\dots,n_i}$, the family $(\varepsilon_{ij})_{ij}$ of error terms shall be iid with components having absolutely continuous cdf as well.

In order for (5.1) being a linear quantile mixed model with respect to the fixed rate $\tau \in (0,1)$, we restrict the error terms (more precisely their cdf) to attain their τ-quantile at 0, hence a convenient choice for the cdf of the error terms is the asymmetric Laplace distribution. More precisely, we assume that $\varepsilon_{ij} \sim \mathrm{AL}(0,1,\tau)$. Eventually, we arrive at the following (parametric) statistical model accompanying our linear quantile mixed model in (5.1)

$$\Big(\mathbf{R}^N, \mathfrak{B}(\mathbf{R}^N), \big(\mathsf{P}_N^{X\beta_\tau+ZU+\varepsilon} \colon \beta_\tau \in \mathbf{R}^p,\ \Psi_\tau \in \mathbf{R}^{q\times q} \text{ spd matrix}\big)\Big). \tag{5.3}$$

Note that our assumptions imply that the Y_i are independent random variables. In other words, the statistical model in (5.3) is the product of the M statistical models on cluster level

$$\Big(\mathbf{R}^{n_i}, \mathfrak{B}(\mathbf{R}^{n_i}), \big(\mathsf{P}^{X_i\beta_\tau+Z_iU_i+\varepsilon_i} \colon \beta_\tau \in \mathbf{R}^p,\ \Psi_\tau \in \mathbf{R}^{q\times q} \text{ spd matrix}\big)\Big). \tag{5.4}$$

Moreover the Y_i are independent given the U_i, i.e. $\mathsf{P}^{Y|U} = \bigotimes_{i=1}^M \mathsf{P}^{Y_i|U_i}$ almost surely. In particular our assumptions guarantee

$$\mathsf{P}^{Y_i|U_i=u_i} = \mathrm{AL}(X_i\beta_\tau + Z_iu_i, \sigma_\tau, \tau), \quad \forall u_i \in \mathbf{R}^{n_i}. \tag{5.5}$$

Therefore, the common probability density of Y and U is given by

$$f^{(Y,U)}(y,u) = f^{Y|U=u}(y)f^U(u) = \prod_{i=1}^M f^{Y_i|U_i=u_i}(y_i \mid \beta_\tau) f^{U_i}(u_i \mid \Psi_\tau),$$

and due to (5.5) for any $y_i = (y_{i1}, \dots, y_{in_i}) \in \mathbf{R}^{n_i}$

$$f^{Y_i|U_i=u_i}(y_i \mid \beta_\tau) = \big(\tau(1-\tau)\big)^{n_i} \exp\Big(-\sum_{j=1}^{n_i} \rho_\tau\big(y_{ij} - x_{ij}^T\beta_\tau - z_{ij}^Tu_i\big)\Big).$$

In the latter formula, $x_{ij} \in \mathbf{R}^p$ (resp. $z_{ij} \in \mathbf{R}^q$) represents the jth row of the design matrix X_i (resp. Z_i) and $\rho_\tau(x) = x\,(\tau - \mathbf{1}^-(x))$ is the loss function of quantile regression. Via

disintegration we obtain the likelihood functions to our statistical model in (5.3)

$$\mathcal{L}(\beta_\tau, \Psi_\tau \,|\, y) = \prod_{i=1}^{M} \mathcal{L}^i(\beta_\tau, \Psi_\tau \,|\, y_i) \tag{5.6}$$

where

$$\mathcal{L}^i(\beta_\tau, \Psi_\tau \,|\, y_i) = \int_{\mathbf{R}^q} \mathrm{d}u_i \, f^{Y_i|U_i=u_i}(y_i \,|\, \beta_\tau) f^{U_i}(u_i \,|\, \Psi_\tau) \tag{5.7}$$

are the likelihood functions to (5.4).

As we already suppressed the τ dependence of the random effects and the error terms, we will continue in this fashion for the parameters β_τ and Ψ_τ in order to ease notation and enhance readability in what follows.

5.2.2 Objective function

In this work we restrict ourselves to the case that the random effects are Gaussian, or more precisely $U_i \sim N(0, \sigma^2 \mathrm{Id}_{q\times q})$. Another choice is given by random effects which are distributed according to the law of a Laplacian (Geraci and Bottai, 2014).

For computational convenience only, we reparametrize the Gaussian density according to the change of variables $\theta = \frac{1}{\sigma^2}$, i.e., we work with

$$f^{U_i}(u_i|\theta) = \sqrt{\frac{\theta^q}{(2\pi)^q}} \exp\Big(-\frac{1}{2}\theta|u_i|^2 \Big). \tag{5.8}$$

Under this assumption, we then study the following (random) objective function

$$\mathbf{R}^p \times \mathbf{R}_{>0} \ni (\delta, \gamma) \mapsto \sum_{i=1}^{M} \log \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i|\gamma) \, f^{Y_i|U_i=u_i}(y_i|\delta), \tag{5.9}$$

where $y \in \mathbf{R}^N$. Note that the objective function is nothing else but the log-likelihood of our linear quantile mixed model (5.3). In other words, the local maximizer of the objective function (in the following always denoted by $\widehat{\beta}_N$ and $\widehat{\theta}_N$) are local maximum-likelihood estimates. In what follows, we will refer to such estimates as regression quantiles and we aim to show that under certain assumptions, there is at least one such sequence which obeys nice asymptotic properties.

To this end, the asymptotic analysis will be carried out on the product space

$$\left(\bigtimes_{N=1}^{\infty} \mathbf{R}^N, \bigotimes_{N=1}^{\infty} \mathfrak{B}\big(\mathbf{R}^N\big), \mathsf{P}^{\beta,\theta} \right),$$

where

$$\mathsf{P}^{\beta,\theta} = \bigotimes_{N=1}^{\infty} \mathsf{P}_N^{\beta,\theta}, \quad \mathsf{P}_N^{\beta,\theta} = \mathsf{P}_N^{X\beta+ZU+\varepsilon}.$$

Strictly speaking, for fixed $N \in \mathbf{N}$, all quantities like the maximum-likelihood estimators $\widehat{\beta}_N$ and $\widehat{\theta}_{M(N)}$ (or the functionals I_N below) are introduced such that they initially live on the measure space $\big(\mathbf{R}^N, \mathfrak{B}\big(\mathbf{R}^N\big)\big)$. But in all of these cases, it is clear how to lift them via coordinate projections to the product space mentioned above so that all of them eventually live on the above product probability space. As this lifting process is so natural, we refrain from incorporating this mere technicality in the notation in order not to clutter our presentation. Instead, we will keep our notation as introduced.

The above mentioned objective function comes along with two technical obstructions. Firstly, our optimization problem is not jointly concave in the parameters and secondly, we have to deal with the singularity at the origin of the objective function of quantile regression. Apart from that, we also have to acknowledge the fact that the very aim of a mixed effects model is to allow for dependencies. All these obstructions together make the problem quite non-standard and we will encounter below how to address them under certain (admittably high-level) restrictions on the design.

5.2.3 Assumptions on the design

Now, let us state and discuss the regularity conditions on the design under which we will carry out the asymptotic analysis. We start with

Assumption 1. We assume for the design matrix linking the random effects U_i with the response Y_i that $z_{ij} \neq 0$ for all $i = 1, \ldots, M$ and all $j = 1, \ldots, n_i$.

This assumption is needed in order to ensure the existence of the first derivative of the log-likelihood with respect to the parameter $\beta \in \mathbf{R}^p$. Remind that regularity with respect to the parameter $\beta \in \mathbf{R}^p$ is *a priori* not obvious due to the singularity of the tilted absolute value function ρ_τ. On the other side, the above assumption for the Z-design together with the disintegration of the random effects already guarantees that the singularity gets restricted to a hyperplane in $\mathbf{R}^q$.

Assumption 2. We assume that

$$M(N) \to \infty, \quad \limsup_{N\to\infty} \max_{i=1,\ldots,M} n_i(N) < \infty.$$

Furthermore, we require that

$$\max_{i=1,\ldots,M} \sum_{j=1}^{n_i} \|x_{ij}\| \in O(N^{\frac{1}{4}}) \quad \text{and} \quad \sum_{i=1}^{M} \left(\sum_{j=1}^{n_i} \|x_{ij}\| \right)^3 \in O(N) \quad \text{as} \quad N \to \infty.$$

The first part of this assumption deals with the precise asymptotic setting for the various sample sizes. The upper bound on the within-cluster sample sizes is strictly speaking not needed for the proof of the main result of this work. However, it necessarily follows from the other conditions in the case that the model features an intercept which is why we preferred to state it explicitly. The latter part concerning the X-design is a suitable generalization of conditions (or slight versions of them) which are commonly used in the case of fixed effects quantile models (Knight, 1998; Knight, 2003; Feng et al., 2011; Koenker, 2005, Sec. 4). As it turns out, the appearing within-cluster summation is crucial. Heuristically, this shall be no surprise as we allow in our model for within-cluster dependencies.

Assumption 3. We assume that there exists a matrix $\overline{B}^{(1)} \in \mathbf{R}^{p\times p}$ and a vector $\overline{B}^{(3)} \in \mathbf{R}^p$ such that

$$\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i} x_{ij}x_{ij}^T \to \overline{B}^{(1)} \quad \text{and} \quad \frac{1}{\sqrt{M}\sqrt{N}}\sum_{i=1}^{M}\sum_{j=1}^{n_i} x_{ij} \to \overline{B}^{(3)} \quad \text{as} \quad N\to\infty.$$

Assumption 4. Furthermore, there exists a matrix $\widetilde{B}^{(1)} \in \mathbf{R}^{p\times p}$, a vector $\widetilde{B}^{(3)} \in \mathbf{R}^p$ and a number $\widetilde{B}^{(2)} \in \mathbf{R}$ such that, as $N\to\infty$,

$$\frac{1}{N}\sum_{i=1}^{M}\sum_{j,k=1}^{n_i} \mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ij})\big|Y_i=\cdot\big]\times\mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ik})\big|Y_i=\cdot\big]\Big]x_{ij}x_{ik}^T \to \widetilde{B}^{(1)},$$

$$\frac{1}{2\sqrt{N}\sqrt{M}}\sum_{i=1}^{M}\sum_{j=1}^{n_i} \mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ij})\big|Y_i=\cdot\big]\times\mathsf{E}\big[U_i^TU_i\big|Y_i=\cdot\big]\Big]x_{ij} \to \widetilde{B}^{(3)}$$

as well as

$$\frac{1}{4M}\sum_{i=1}^{M} \mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[U_i^TU_i\big|Y_i=\cdot\big]^2\Big] \to \widetilde{B}^{(2)}.$$

The last two assumptions represent further moment conditions on the design. Note that in the context of a mixed effects model the conditional distribution given the response apparently plays a prominent role. In particular, the above assumptions not only impose restrictions on the X-design but also on the Z-design which links the random effects with the response. Note also that our moment assumptions for the design contain the assumptions typically employed in fixed effects quantile regression models with error terms distributed according to the law of an asymmetric Laplacian (Koenker, 2005, Sec. 4).

Assumption 5. The block matrix $\mathbf{B} \in \mathbf{R}^{(p+1)\times(p+1)}$ given by

$$\mathbf{B} = \begin{pmatrix} B^{(1)} & B^{(3)} \\ \{B^{(3)}\}^T & B^{(2)} \end{pmatrix} = \begin{pmatrix} \widetilde{B}^{(1)} - \tau^2\overline{B}^{(1)} & \widetilde{B}^{(3)} - \frac{q}{2\theta}\tau\overline{B}^{(3)} \\ \big(\widetilde{B}^{(3)} - \frac{q}{2\theta}\tau\overline{B}^{(3)}\big)^\intercal & \widetilde{B}^{(2)} - \frac{q^2}{4\theta^2} \end{pmatrix}$$

is assumed to be positive definite.

The matrix $\mathbf{B}$ encodes the building block for the limit covariance matrix. In order to rule out certain degenerate situations in the limit we impose the above restriction on the eigenvalues of the limit matrix $\mathbf{B}$. Furthermore, in the special case of a model with only fixed effects, i.e., $M = N$, $n_i = 1$ and $\mathsf{P}^{U_i} = \delta_0$, one obtains due to our model assumption on the error term that

$$\mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ij})\big|Y_i = \cdot\big] \times \mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ik})\big|Y_i = \cdot\big]\Big] = \tau\delta_{jk}.$$

In particular, it follows in this case that

$$B^{(1)} = \tau(1-\tau) \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T.$$

Therefore, in our work we recover the well-known result from the asymptotic theory for linear quantile regression models with asymmetric Laplacian error terms (Koenker, 2005, Sec. 4). We conclude with a "compactness-type" assumption.

Assumption 6. Let $\mathfrak{K}(\beta,\theta)$ denote the following compact subset of $\mathbf{R}^p \times \mathbf{R}_{>0}$:

$$\Big\{(\widetilde{\beta},\widetilde{\theta}) \in \mathbf{R}^{p+1} \colon \|\widetilde{\beta}-\beta\| \leq 1 \text{ and } |\widetilde{\theta}-\theta| \leq 1 \wedge \frac{\theta}{2}\Big\}.$$

We then assume that

$$\sup_{N\in\mathbf{N}} \sup_{i=1,\ldots,M} \sup_{(\widetilde{\beta},\widetilde{\theta})\in\mathfrak{K}(\beta,\theta)} \Big\| \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=\cdot}(\mathrm{d}u_i|\widetilde{\beta},\widetilde{\theta})\,|u_i|^4 \Big\|_{L^\infty(\mathbf{R}^{n_i})} \lesssim 1.$$

This assumption together with the second one represent sufficient conditions in order to guarantee uniform convergence of the remainder term originating from a kind of weak second-order Taylor approximation of our objective function (cf. condition (5.25) in the formulation of Lemma 5.4.1). It obviously represents another moment condition for the conditional distribution given the response and therefore again incorporates a restriction on the X-design as well as Z-design.

5.2.4 Main result

Our main theoretical result concerns the asymptotic behavior of regression quantiles in linear quantile mixed effects models and reads as follows.

Theorem 5.2.1. *Fix parameters $\beta \in \mathbf{R}^p$, $\theta > 0$, and consider an associated sequence of linear quantile mixed effects models. In particular, let the assumptions and notation of Sections 5.2.1–5.2.3 be in place. In this situation, there exists at least one sequence of measurable functions*

$$\widehat{\beta}_N \colon \mathbf{R}^N \to \mathbf{R}^p, \quad \widehat{\theta}_M \colon \mathbf{R}^N \to \mathbf{R}_{>0}$$

such that the following two properties hold true:

i) With probability reaching one under $\mathsf{P}^{\beta,\theta}$ *as* $N \to \infty$, $\big(\widehat{\beta}_N, \widehat{\theta}_M\big)$ *is a local maximizer of our* log*-likelihood.*

ii) The appropriately centered and rescaled sequence

$$\big(\sqrt{N}(\widehat{\beta}_N - \beta)^T, \sqrt{M}(\widehat{\theta}_M - \theta)\big)^T$$

converges under $\mathsf{P}^{\beta,\theta}$ *in distribution as* $N \to \infty$*. The limit distribution is a centered Gaussian with covariance structure* $\mathbf{B}^{-1}$*. Moreover, the following linear representation with respect to* $\mathsf{P}^{\beta,\theta}$ *holds true*

$$\big(\sqrt{N}(\widehat{\beta}_N - \beta)^T, \sqrt{M}(\widehat{\theta}_M - \theta)\big)^T = \mathbf{B}_N^{-1}\nabla_{(\beta,\theta)}\ell(\beta,\theta|\cdot) + o_p(1),$$

where

$$\mathbf{B}_N = \operatorname{diag}\big(\sqrt{N}\mathrm{Id}_{p\times p}, \sqrt{M}\big)\mathbf{B} \in \mathbf{R}^{(p+1)\times(p+1)}.$$

Remark 1. It is a well-known fact from the asymptotic theory of fixed effects quantile regression models that the particular form of the asymptotic distribution heavily depends on the behaviour of the distribution of the response around the considered quantile (Koenker, 2005). In our case, we have *due to our model assumption for the error terms* that

$$f^{Y_{ij}|U_i=u_i}\big(Q^{Y_{ij}|U_i=u_i}(\beta)\big) = f^{Y_{ij}|U_i=u_i}\big(x_{ij}^T\beta + z_{ij}^T u_i\big) = \tau(1-\tau),$$

which is uniformly bounded away from zero and from above. In particular, one can expect that in the mixed effects model as described above the limit distribution is given by a Gaussian.

5.3 Computation of derivatives of the log-likelihood

In this section, we compute the derivatives of the log-likelihood

$$\ell(\beta,\theta|y) = \sum_{i=1}^{M} \log \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i|\theta)\, f^{Y_i|U_i=u_i}(y_i|\beta).$$

We begin with derivatives with respect to the parameter $\beta \in \mathbf{R}^p$. To this end, we will suppress the dependence on the parameter $\theta > 0$ in the proofs of the following two technical lemmas.

Lemma 5.3.1. *We have* $\ell(\cdot,\theta|y) \in \mathcal{C}^1(\mathbf{R}^p)$ *for all* $y \in \mathbf{R}^N$*. Furthermore, it holds*

$$\nabla_\beta \ell(\beta,\theta|y) = \sum_{i=1}^{M}\sum_{j=1}^{n_i} \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta)\Big\{\tau - \mathbf{1}^-\big(y_{ij} - x_{ij}^T\beta - z_{ij}^T u_i\big)\Big\}x_{ij}.$$

Proof. Let e_l be the l-th standard basis vector of $\mathbf{R}^p$ and consider $h \neq 0$. In a first step, we

study the differentiability of the function

$$g(\beta|y) = f^{Y_i}(y_i|\beta) = \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i)\, f^{Y_i|U_i=u_i}(y_i|\beta).$$

For this, first note that we have the bound

$$\left| f^{Y_i|U_i=u_i}(y_i|\beta + he_l) - f^{Y_i|U_i=u_i}(y_i|\beta) \right| \lesssim |h| \sum_{j=1}^{n_i} \|x_{ij}\|. \tag{5.10}$$

This is true as $x \mapsto \exp(-x)$ is globally Lipschitz on $\mathbf{R}_{\geq 0}$ and ρ_τ is globally Lipschitz on $\mathbf{R}$. In particular, the difference quotient

$$\left| \frac{f^{Y_i|U_i=u_i}(y_i|\beta + he_l) - f^{Y_i|U_i=u_i}(y_i|\beta)}{h} \right|$$

is bounded from above by an P^{U_i}-integrable function. Therefore, it suffices to show that the difference quotient converges P^{U_i}-almost surely. For this, we take a look at the linear equation

$$y_{ij} - x_{ij}^T \beta = z_{ij}^T u_i.$$

(Remind that $\beta \in \mathbf{R}^p$ is fixed for the moment.) The set of $u_i \in \mathbf{R}^q$ satisfying this equation is an affine subspace of dimension $q-1$ in $\mathbf{R}^q$ if $z_{ij} \neq 0$. The latter, however, is forced to hold by means of Assumption 1. Hence, the set of $u_i \in \mathbf{R}^q$ satisfying the above equation is a set with Lebesgue measure zero in $\mathbf{R}^q$. On the other side, for $u_i \in \mathbf{R}^q$ outside of this set of measure zero we can differentiate ρ_τ with respect to $\beta \in \mathbf{R}^p$ as we are away from the singularity. A simple application of the chain rule then reveals that

$$\begin{aligned} &\left| \frac{f^{Y_i|U_i=u_i}(y_i|\beta + he_l) - f^{Y_i|U_i=u_i}(y_i|\beta)}{h} \right| \\ &\to f^{Y_i|U_i=u_i}(y_i|\beta) \Big\{ \tau - \mathbf{1}^-\big(y_{ij} - x_{ij}^T\beta - z_{ij}^T u_i\big) \Big\} x_{ij}(l) \end{aligned}$$

as $h \to 0$ outside a set of measure zero in $\mathbf{R}^q$. All in all, we infer by means of Lebesgue's dominated convergence theorem and from our assumption that P^{U_i} admits a density that

$$\nabla_\beta g(\beta|y) = \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i)\, f^{Y_i|U_i=u_i}(y_i|\beta) \Big\{ \tau - \mathbf{1}^-\big(y_{ij} - x_{ij}^T\beta - z_{ij}^T u_i\big) \Big\} x_{ij}$$

for all $y \in \mathbf{R}^N$ and $\beta \in \mathbf{R}^p$. Note that this function is continuous in $\beta \in \mathbf{R}^p$. Hence, all what remains to do is to exploit the chain rule in order to calculate the derivative of the log-likelihood. □

Remark 2. Note that the disintegration of the random effects in combination with Assumption 1 has a smoothing effect in the sense that the non-differentiability of the objective function of quantile regression becomes negligible.

Lemma 5.3.2. *Fix $\beta \in \mathbf{R}^p$. Then, the function $\delta \mapsto \mathsf{E}^{\beta,\theta}\big[\nabla \ell(\delta, \theta|\cdot)\big]$ is differentiable at the point $\delta_0 = \beta$ and it holds*

$$\begin{aligned}\nabla_\delta \mathsf{E}^{\beta,\theta}\big[\nabla \ell(\delta, \theta|\cdot)\big]\big|_{\delta=\beta} &= -\tau(1-\tau)\sum_{i=1}^{M}\sum_{j=1}^{n_i} x_{ij}x_{ij}^T \\ &+ \sum_{i=1}^{M}\sum_{j,k=1}^{n_i} \mathsf{E}^{\beta,\theta}\Big[\mathbf{Cov}\big[\tau - \mathbf{1}^-(\varepsilon_{ij}); \tau - \mathbf{1}^-(\varepsilon_{ik})\big|Y_i = \cdot\big]\Big] x_{ij}x_{ik}^T.\end{aligned}$$

Remark 3. Note that we need another integration (this time we disintegrate the response) in order to obtain higher regularity with respect to the model parameter $\beta \in \mathbf{R}^p$.

Proof. One of the main technical parts of the proof consists of deriving a slightly refined version of the bound in (5.10). So, let us start with this aim. A simple application of the fundamental theorem of calculus yields (notation as above)

$$\begin{aligned}&\big|f^{Y_i|U_i=u_i}(y_i|\beta + he_l) - f^{Y_i|U_i=u_i}(y_i|\beta)\big| \\ &\lesssim f^{Y_i|U_i=u_i}(y_i|\beta)\int_0^1 \mathrm{d}t\, \exp\big(-t(x^i - x_0^i)\big)(x^i - x_0^i),\end{aligned}$$

where we introduced the shorthands

$$x_0^i = \sum_{j=1}^{n_i}\rho_\tau\big(y_{ij} - x_{ij}^T\beta - z_{ij}^T u_i\big), \quad x^i = \sum_{j=1}^{n_i}\rho_\tau\big(y_{ij} - x_{ij}^T(\beta + he_l) - z_{ij}^T u_i\big).$$

Now, we will again make use of the Lipschitz property of ρ_τ, i.e., there is a constant $c = c(\tau) > 0$ such that

$$|x^i - x_0^i| \le c(\tau)|h|\sum_{j=1}^{n_i}\|x_{ij}\|.$$

Hence, by means of the monotonicity of the exponential we obtain the bound

$$\begin{aligned}&\big|f^{Y_i|U_i=u_i}(y_i|\beta + he_l) - f^{Y_i|U_i=u_i}(y_i|\beta)\big| \\ &\lesssim f^{Y_i|U_i=u_i}(y_i|\beta)e^{c(\tau)|h|\sum_{j=1}^{n_i}\|x_{ij}\|}\sum_{j=1}^{n_i}\|x_{ij}\||h|,\end{aligned} \tag{5.11}$$

and the proportionality constant implicit in this bound depends only on τ. (Of course, we can restrict ourselves to the situation $|h| \le 1$.) As it turns out, this bound will be sufficient to calculate the desired derivative.

To this end, note first that

$$\mathsf{E}^{\beta,\theta}\big[\nabla \ell(\delta|\cdot)\big]$$

$$= \sum_{i=1}^{M} \sum_{j=1}^{n_i} \int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta) \int_{\mathbf{R}^q} \underbrace{\mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\delta) \Big\{\tau - \mathbf{1}^-\big(y_{ij} - x_{ij}^T\delta - z_{ij}^T u_i\big)\Big\} x_{ij}}_{=g^{ij}(\delta)}.$$

In order to streamline the exposition furthermore, we introduce the shorthands

$$e'_{ij}(\delta) = y_{ij} - Q_{ij}(\delta), \quad Q_{ij}(\delta) = x_{ij}^T\delta + z_{ij}^T u_i.$$

As above, we will now study the difference quotient under the integral around the point $\delta_0 = \beta$. For this, we use the fact that we can write

$$\begin{aligned} g^{ij}(\delta + he_l) - g^{ij}(\delta) = &-f^{U_i|Y_i=y_i}(u_i|\delta)\Big\{\mathbf{1}^-\big(e'_{ij}(\delta + he_l)\big) - \mathbf{1}^-\big(e'_{ij}(\delta)\big)\Big\} \\ &+ \Big\{f^{U_i|Y_i=y_i}(u_i|\delta + he_l) - f^{U_i|Y_i=y_i}(u_i|\delta)\Big\} \cdot \Big\{\tau - \mathbf{1}^-\big(e'_{ij}(\delta + he_l)\big)\Big\}. \end{aligned}$$

The first term can be dealt with directly. More precisely, we compute (the letter F is, as usual, referring to the cumulative distribution function)

$$\begin{aligned} &\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta) \int_{\mathbf{R}^q} \mathrm{d}u_i \, f^{U_i|Y_i=y_i}(u_i|\delta_0)\Big\{\mathbf{1}^-\big(e'_{ij}(\delta_0 + he_l)\big) - \mathbf{1}^-\big(e'_{ij}(\delta_0)\big)\Big\} \\ &= \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i) \Big\{F^{Y_{ij}|U_i=u_i}\big(Q_{ij}(\delta_0) + x_{ij}^T he_l|\delta_0\big) - F^{Y_{ij}|U_i=u_i}\big(Q_{ij}(\delta_0)|\delta_0\big)\Big\}. \end{aligned}$$

Hence, if we divide by $h \neq 0$ we immediately obtain due to the usual differentiation under the integral sign that the latter term converges to

$$\int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i) \underbrace{f^{Y_{ij}|U_i=u_i}\big(Q_{ij}(\delta_0)|\delta_0\big)}_{=\tau(1-\tau)} x_{ij}(l) = \tau(1-\tau)x_{ij}(l) \tag{5.12}$$

as $h \to 0$. Note at this point, that this argument heavily relies on the fact that we integrated over the response and not just over the random effects.

For the second term, we exploit our bound in (5.11) and what we have achieved throughout the proof of Lemma 5.3.1. In fact, what we will do is just a careful application of the quotient rule under the integral sign. This can best be done by further decomposing the second term as follows, using in particular the well-know relations for conditional densities,

$$\begin{aligned} &\int_{\mathbf{R}^q} \mathrm{d}u_i \Big\{f^{U_i|Y_i=y_i}(u_i|\delta_0 + he_l) - f^{U_i|Y_i=y_i}(u_i|\delta_0)\Big\} \cdot \Big\{\tau - \mathbf{1}^-\big(e'_{ij}(\delta_0 + he_l)\big)\Big\} \\ &= \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i) \, \frac{f^{Y_i|U_i=u_i}(y_i|\delta_0 + he_l) - f^{Y_i|U_i=u_i}(y_i|\delta_0)}{f^{Y_i}(y_i|\delta_0)} \Big\{\tau - \mathbf{1}^-\big(e'_{ij}(\delta_0 + he_l)\big)\Big\} \\ &\quad + \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i) \, \left\{\frac{1}{f^{Y_i}(y_i|\delta_0 + he_l)} - \frac{1}{f^{Y_i}(y_i|\delta_0)}\right\} f^{Y_i|U_i=u_i}(y_i|\delta_0 + he_l) \\ &\qquad\qquad \times \Big\{\tau - \mathbf{1}^-\big(e'_{ij}(\delta_0 + he_l)\big)\Big\}. \end{aligned}$$

Again, we treat both terms in this sum separately. If we divide the first one by h, the resulting

term admits due to (5.11) the bound

$$\left| \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i) \frac{f^{Y_i|U_i=u_i}(y_i|\delta_0 + he_l) - f^{Y_i|U_i=u_i}(y_i|\delta_0)}{h \cdot f^{Y_i}(y_i|\delta_0)} \left\{ \tau - \mathbf{1}^-\big(e'_{ij}(\delta_0 + he_l)\big) \right\} \right|$$
$$\lesssim \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i) \, \frac{f^{Y_i|U_i=u_i}(y_i|\delta_0)}{f^{Y_i}(y_i|\delta_0)}$$

over $|h| \leq 1$. But the latter function is integrable over the domain $\mathbf{R}^{n_i}$ with respect to the measure P^{Y_i}. Thus, if we integrate over $\mathbf{R}^{n_i}$ with respect to the measure P^{Y_i} and let $h \to 0$ we obtain by dominated convergence the following limit for the first term:

$$\sum_{k=1}^{n_i} \mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\{\tau - \mathbf{1}^-(\varepsilon_{ij})\} \times \{\tau - \mathbf{1}^-(\varepsilon_{ik})\} \big| Y_i = \cdot \big] \Big] x_{ik}(l). \tag{5.13}$$

Furthermore, by an application of the mean-value inequality the second term can be bounded over $|h| \leq 1$ by

$$\sup_{\delta \in B_1(\delta_0)} \left| \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\delta) \right|^2 \leq 1$$

times a proportionality constant. Thus, another application of the dominated convergence theorem of Lebesgue yields as the limit for the second term (after integrating and dividing by h, of course) the quantity

$$-\sum_{k=1}^{n_i} \mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\tau - \mathbf{1}^-(\varepsilon_{ij}) \big| Y_i = \cdot \big] \times \mathsf{E}\big[\tau - \mathbf{1}^-(\varepsilon_{ik}) \big| Y_i = \cdot \big] \Big] x_{ik}(l). \tag{5.14}$$

Therefore, the claim follows. □

Remark 4. A straightforward computation shows that the following relation

$$\mathsf{E}^{\beta,\theta}\Big[\big(\nabla \ell(\beta, \theta|\cdot) \big)^{\otimes 2} \Big] = -\nabla \mathsf{E}^{\beta,\theta}\Big[\nabla \ell(\beta, \theta|\cdot) \Big] \tag{5.15}$$

holds true. Indeed, one immediately recognizes that the left hand side is given by the quantity

$$\sum_{i=1}^{M} \sum_{j,k=1}^{n_i} \mathbf{Cov}^{\beta,\theta}\Big[\mathsf{E}\big[\tau - \mathbf{1}^-(\varepsilon_{ij}) \big| Y_i = \cdot \big]; \mathsf{E}\big[\tau - \mathbf{1}^-(\varepsilon_{ik}) \big| Y_i = \cdot \big] \Big]$$
$$= \sum_{i=1}^{M} \sum_{j,k=1}^{n_i} \mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\tau - \mathbf{1}^-(\varepsilon_{ij}) \big| Y_i = \cdot \big] \times \mathsf{E}\big[\tau - \mathbf{1}^-(\varepsilon_{ik}) \big| Y_i = \cdot \big] \Big],$$

since $\mathsf{E}[\tau - \mathbf{1}^-(\varepsilon_{ij})|Y_i = \cdot]$ is a centered random variable under $\mathsf{P}^{\beta,\theta}$. On the other side, the term in (5.13) actually reduces to

$$\sum_{k=1}^{n_i} \tau(1-\tau) x_{ik}(l),$$

i.e., it cancels the term appearing in (5.12) and we thus obtain (5.15). Due to the low regularity of the objective function of quantile regression, this is the suitable generalization of the well-known relation in maximum-likelihood theory

$$\mathsf{E}^{\beta,\theta}\Big[\big(\nabla\ell(\beta,\theta|\cdot)\big)^2\Big] = -\mathsf{E}^{\beta,\theta}\Big[\nabla^2\ell(\beta,\theta|\cdot)\Big], \tag{5.16}$$

which only holds in more regular cases.

Let us now state the derivatives with respect to the parameter $\theta > 0$. We omit the proof as it simply boils down to the calculation of the derivatives of a Gaussian with respect to the precision parameter.

Lemma 5.3.3. *We have $\theta \mapsto \ell(\beta,\cdot|y) \in \mathcal{C}^\infty(\mathbf{R}_{>0})$ for all $y \in \mathbf{R}^N$. Moreover, one computes*

$$\frac{\partial}{\partial\theta}\ell(\beta,\theta|y) = \sum_{i=1}^{M}\mathsf{E}\bigg[\frac{q}{2\theta} - \frac{1}{2}U_i^TU_i\bigg|Y_i = y_i\bigg]$$

and

$$\frac{\partial^2}{\partial\theta^2}\ell(\beta,\theta|y) = -M\frac{q}{2\theta^2} + \sum_{i=1}^{M}\mathbf{Var}\bigg[\frac{q}{2\theta} - \frac{1}{2}U_i^TU_i\bigg|Y_i = y_i\bigg].$$

We conclude this section with the result for the mixed second-order derivatives of the log-likelihood. Again, we simply state it.

Lemma 5.3.4. *It holds that $\nabla_\beta\ell(\beta,\theta|y) \in \mathcal{C}^\infty(\mathbf{R}_{>0})$ and $\frac{\partial}{\partial\theta}\ell(\beta,\theta|y) \in \mathcal{C}^1(\mathbf{R}^p)$ for every $y \in \mathbf{R}^N$. In addition,*

$$\begin{aligned}\frac{\partial}{\partial\theta}\nabla_\beta\ell(\beta,\theta|y) &= \nabla_\beta\frac{\partial}{\partial\theta}\ell(\beta,\theta|y)\\ &= \sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{Cov}\bigg[\tau - \mathbf{1}^-(\varepsilon_{ij}); \frac{q}{2\theta} - \frac{1}{2}U_i^TU_i\bigg|Y_i = y_i\bigg]x_{ij}.\end{aligned}$$

Remark 5. As in the remark above, one can check that the following two relations hold true

$$\mathsf{E}^{\beta,\theta}\bigg[\Big(\frac{\partial}{\partial\theta}\ell(\beta,\theta|\cdot)\Big)^2\bigg] = -\mathsf{E}^{\beta,\theta}\bigg[\frac{\partial^2}{\partial\theta^2}\ell(\beta,\theta|\cdot)\bigg]$$

and

$$\mathsf{E}^{\beta,\theta}\bigg[\frac{\partial}{\partial\theta}\ell(\beta,\theta|\cdot)\nabla_\beta\ell(\beta,\theta|\cdot)\bigg] = -\mathsf{E}^{\beta,\theta}\bigg[\frac{\partial}{\partial\theta}\nabla_\beta\ell(\beta,\theta|\cdot)\bigg].$$

5.4 Proof of Theorem 5.2.1

5.4.1 Outline of the strategy

Before we give the proof of our main result, we want to comment on the idea and structure of the proof. The main point consists of studying the following sequence of random functionals:

$$\begin{aligned} I_N(\delta,\gamma|y) = &\sum_{i=1}^{M} \log \int_{\mathbf{R}^q} \mathsf{P}^{U_i}\big(\mathrm{d}u_i\big|\theta + M^{-\frac{1}{2}}\gamma\big)\, f^{Y_i|U_i=u_i}\big(y_i\big|\beta + N^{-\frac{1}{2}}\delta\big) \\ &- \sum_{i=1}^{M} \log \int_{\mathbf{R}^q} \mathsf{P}^{U_i}(\mathrm{d}u_i|\theta)\, f^{Y_i|U_i=u_i}(y_i|\beta). \end{aligned}$$

As always, the idea behind this is the simple observation that if $(\widehat{\beta}_N, \widehat{\theta}_M)$ is a local maximizer of our objective function (5.9) then

$$\big(\widehat{\delta}_N, \widehat{\gamma}_M\big) = \big(\sqrt{N}(\widehat{\beta}_N - \beta), \sqrt{M}(\widehat{\theta}_M - \theta)\big)$$

is a local maximizer of the functional I_N (and vice versa). Furthermore, the very definition of the functional I_N also motivates to expand it via

$$I_N(\delta,\gamma|y) = \underbrace{\frac{1}{\sqrt{N}}\nabla_\beta \ell(\beta,\theta|y)^T}_{=A_N^{(1)}(\beta,\theta)}\,\delta + \underbrace{\frac{1}{\sqrt{M}}\frac{\partial}{\partial\theta}\ell(\beta,\theta|y)}_{=A_N^{(2)}(\beta,\theta)}\,\gamma + R_N^{(1)}(\delta,\gamma|y). \tag{5.17}$$

Based on our assumptions, it is rather straightforward to check that the sequence $\big(A_N^{(1)}(\beta,\theta)^T, A_N^{(2)}(\beta,\theta)\big)^T$ converges in distribution under $\mathsf{P}^{\beta,\theta}$ as $N \to \infty$. The limit distribution is moreover given by a centered Gaussian with covariance matrix $\mathbf{B} \in \mathbf{R}^{(p+1)\times(p+1)}$.

In a second step, we would like to expand the remainder term $R_N^{(1)}(\delta,\gamma|y)$. At this point, we have to be a little bit careful as the log-likelihood is not twice continuously differentiable in the classical sense with respect to the model parameter $\beta \in \mathbf{R}^p$. But as we have seen in the previous section, our log-likelihood admits a second-order approximation in a weaker sense. More precisely, we may expand the expected value of the first-order remainder $R_N^{(1)}(\delta,\gamma)$ in the sense

$$\begin{aligned} &\mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big] \\ &= -\frac{1}{2}\delta^T B_N^{(1)}(\beta,\theta)\delta - \frac{1}{2}\gamma B_N^{(2)}(\beta,\theta)\gamma - \gamma B_N^{(3)}(\beta,\theta)\delta + R_N^{(2)}(\delta,\gamma|y), \end{aligned}$$

where

$$\begin{aligned} B_N^{(1)} = -\frac{1}{N}\nabla\mathsf{E}^{\beta,\theta}\Big[\nabla\ell(\beta,\theta|\cdot)\Big], \quad B_N^{(2)}(\beta,\theta) = -\frac{1}{M}\mathsf{E}^{\beta,\theta}\bigg[\frac{\partial^2}{\partial\theta^2}\ell(\beta,\theta|\cdot)\bigg], \\ B_N^{(3)}(\beta,\theta) = -\frac{1}{\sqrt{NM}}\mathsf{E}^{\beta,\theta}\bigg[\frac{\partial}{\partial\theta}\nabla\ell(\beta,\theta|\cdot)\bigg]. \end{aligned} \tag{5.18}$$

Adding and subtracting the expected value of $R_N^{(1)}(\delta,\gamma)$ in (5.17), and then inserting the previous expansion therefore yields the representation

$$I_N(\delta,\gamma|y) = A_N^{(1)}(\beta,\theta)\delta + A_N^{(2)}(\beta,\theta)\theta \tag{5.19}$$

$$- \frac{1}{2}\delta^T B_N^{(1)}(\beta,\theta)\delta - \frac{1}{2}\gamma B_N^{(2)}(\beta,\theta)\gamma - \gamma B_N^{(3)}(\beta,\theta)\delta \tag{5.20}$$

$$+ R_N^{(2)}(\delta,\gamma|y) + \Big(R_N^{(1)}(\delta,\gamma|y) - \mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big]\Big). \tag{5.21}$$

Note next that the combination of Assumption 3 and Assumption 4 implies that $B_N^{(i)}(\beta,\theta)$ converges to $B^{(i)}(\beta,\theta)$ as we let $N \to \infty$. Let us write $B_N(\beta,\theta)$ (resp. $\mathbf{B}$) for the associated block matrix

$$B_N(\beta,\theta) = \begin{pmatrix} B_N^{(1)}(\beta,\theta) & B_N^{(3)}(\beta,\theta) \\ B_N^{(3)}(\beta,\theta)^T & B_N^{(2)}(\beta,\theta) \end{pmatrix} \in \mathbf{R}^{(p+1)\times(p+1)}.$$

Note that $\mathbf{B}$ is positive definite by Assumption 5. Let us also introduce the shorthands

$$\Xi = (\delta^T,\gamma)^T \in \mathbf{R}^{p+1}, \quad A_N(\beta,\theta) = \big(A_N^{(1)}(\beta,\theta)^T, A_N^{(2)}(\beta,\theta)\big)^T \in \mathbf{R}^{p+1}.$$

With all of this notation at hand, we can then write in view of (5.19) our second-order expansion by also adding and subtracting $\frac{1}{2}\Xi^T\mathbf{B}\Xi$ as follows

$$I_N(\delta,\gamma|y) = \Xi^T A_N(\beta,\theta) - \frac{1}{2}\Xi^T\mathbf{B}\Xi + R_N^{(3)}(\delta,\gamma|y), \tag{5.22}$$

where $R_N^{(3)}(\delta,\gamma|y)$ denotes some suitable error term which is made more precise in the proof below.

We next observe that the properties of the matrix $\mathbf{B}$ ensure that $\Xi_N^* = (\delta_N^*,\gamma_M^*) = \mathbf{B}^{-1}A_N(\beta,\theta)$ is the unique global maximizer of the strictly concave and quadratic functional

$$\Xi \mapsto \Xi^T A_N(\beta,\theta) - \frac{1}{2}\Xi^T\mathbf{B}\Xi.$$

Note that by the previous arguments Ξ_N^* converges in distribution under $\mathsf{P}^{\beta,\theta}$ as $N \to \infty$, and the corresponding limit distribution is given by a centered Gaussian with covariance structure $\mathbf{B}^{-1}$.

As we have already noted, our maximization problem is not concave in the variable $\Xi \in \mathbf{R}^{p+1}$. Therefore, we cannot directly make use of the well-known *convexity argument* (Hjort and Pollard, 1993; Geyer, 1996; Knight, 1998) in order to conclude the proof. Furthermore, as we model within-cluster dependencies for the response one is tempted to base the remaining analysis on the results of Weiss (Weiss, 1971; Weiss, 1973). For instance, this approach was successfully applied to the asymptotic analysis of linear mixed models for the mean (Pinheiro, 1994). But due to the insufficient regularity of our objective function, we are not exactly in the

situation of the results in Weiss (1971) and Weiss (1973).

On the other side, what all of the above mentioned methods have in common is that one wants to establish the existence of a local maximizer of the functional $I_N(\delta, \gamma|y)$ (at least with probability approaching one as $N \to \infty$). In addition, the procedure shall come along with a probabilistic bound on how far this local maximizer stays away from the Ξ_N^*'s. If this bound is sufficiently sharp, then one can typically conclude the asymptotic normality of these local maximizer from the corresponding statement for the Ξ_N^*'s. Furthermore, this argument also typically yields the asserted linear representation. In order to make this procedure work in our situation, it will be key to establish *uniform* convergence of the remainder. As it turns out in the proof, this is precisely the point of having the "compactness-type" Assumption 6 as a substitute for being outside of the scope of standard arguments.

Let us make this more precise. To this end, we first re-expand around the point Ξ_N^* as follows:

$$I_N(\delta, \gamma|y) = H_N(\Xi_N^*) + R_N^{(3)}(\delta, \gamma|y) - \big(\Xi - \Xi_N^*\big)^T \mathbf{B} \Xi_N^* - \frac{1}{2}\big(\Xi - \Xi_N^*\big)^T \mathbf{B} \big(\Xi - \Xi_N^*\big).$$

The first term appearing here is of no particular importance for us as it only depends on the value of Ξ_N^*. (The precise form of the remainder term will be specified later on.) In particular, we obtain

$$\begin{aligned}
&\ell(\beta + N^{-\frac{1}{2}}\delta_N^*, \theta + M^{-\frac{1}{2}}\gamma_M^*|y) - \ell(\beta + N^{-\frac{1}{2}}\delta, \theta + M^{-\frac{1}{2}}\gamma|y) \\
&\quad = R_N^{(3)}(\delta_N^*, \gamma_M^*|y) - R_N^{(3)}(\delta, \gamma|y) + \big(\Xi - \Xi_N^*\big)^T \mathbf{B} \Xi_N^* + \frac{1}{2}\big(\Xi - \Xi_N^*\big)^T \mathbf{B} \big(\Xi - \Xi_N^*\big).
\end{aligned}$$

Furthermore, consider the sets

$$C^*(N, \epsilon) = \big\{\Xi \in \mathbf{R}^{p+1} \colon \|\Xi - \Xi_N^*\|_{\mathbf{B}} = \epsilon\big\}.$$

As the sequence $(\Xi_N^*)_{N \in \mathbf{N}}$ is tight, we infer that for every $\epsilon_N \to 0$ it holds

$$\max_{\Xi \in C^*(N, |\epsilon_N|)} \big|\big(\Xi - \Xi_N^*\big)^T \mathbf{B} \Xi_N^*\big| \to 0$$

in probability under $\mathsf{P}^{\beta,\theta}$ as $N \to \infty$. Now, consider a sequence $(\kappa_N)_{N \in \mathbf{N}}$ with the properties

$$\kappa_N \to \infty \quad \text{and} \quad \frac{\kappa_N}{\sqrt{M}} \to 0 \quad \text{as} \quad N \to \infty. \tag{5.23}$$

For all what follows, we will assume that $N \in \mathbf{N}$ is already sufficiently large such that it holds

$$-(1 \wedge \theta_-) \leq \frac{\kappa_N}{\sqrt{M}} \leq 1 \wedge \theta_- \quad \text{with } \theta_- := \theta/2. \tag{5.24}$$

As in Weiss (1971, Sec. 4), one can then establish the following important lemma.

Lemma 5.4.1. *Consider the situation described as above and assume that*

$$\sup_{\|\Xi\| \leq \kappa_N} \left| R_N^{(3)}(\delta, \gamma | \cdot) \right| \to 0 \tag{5.25}$$

in probability under $\mathsf{P}^{\beta,\theta}$ *as* $N \to \infty$*. Then, there exists a zero sequence* $(\epsilon_N)_{N \in \mathbf{N}}$ *and at least one sequence* $\widehat{\Xi}_N = (\widehat{\delta}_N, \widehat{\gamma}_M) \in \mathbf{R}^{p+1}$ *such that*

i) with probability reaching one under $\mathsf{P}^{\beta,\theta}$ *as* $N \to \infty$*,*

$$(\widehat{\beta}_N, \widehat{\theta}_N) = (\beta, \theta) + (N^{-\frac{1}{2}} \widehat{\delta}_N, M^{-\frac{1}{2}} \widehat{\gamma}_M)$$

is a local maximizer of our log*-likelihood* (5.9)*, and*

ii) the probability $\mathsf{P}^{\beta,\theta}\big(\|\widehat{\Xi}_N - \Xi_N^*\|_{\mathbf{B}} < \epsilon_N\big)$ *reaches one as* $N \to \infty$*.*

Here, we denoted by $\| \cdot \|_{\mathbf{B}}$ *the norm in* $\mathbf{R}^{p+1}$ *which is induced by the positive definite matrix* $\mathbf{B} \in \mathbf{R}^{(p+1) \times (p+1)}$*. In particular, we obtain that*

$$\big(\sqrt{N}(\widehat{\beta}_N - \beta)^T, \sqrt{M}(\widehat{\theta}_M - \theta)\big)^T \to \mathcal{N}(0, \mathbf{B}^{-1})$$

under $\mathsf{P}^{\beta,\theta}$ *as* $N \to \infty$*.*

Proof. This is just a matter of adapting the arguments in Weiss (1971, Sec. 4) to our notation. □

Therefore, all what remains to do is to carry out the details of the above sketch for the proof and to check the uniformity condition (5.25) for the remainder term. In the course of doing this, we will also specify the sequence $(\kappa_N)_{N \in \mathbf{N}}$.

5.4.2 Proof of Theorem 5.2.1 (Asymptotic normality of regression quantiles in a mixed effects model)

We conduct the proof in several steps and start with

Step 1 (First-order approximation): As already suggested in the preceding discussion, we can write

$$I_N(\delta, \gamma | y) = \Xi^T A_N(\beta, \theta) + R_N^{(1)}(\delta, \gamma | y)$$

where we recall that $A_N(\beta, \theta)$ was defined in (5.17) as the suitably rescaled gradient of the log-likelihood. In other words, it holds by means of Lemma 5.3.1 and Lemma 5.3.3, respec-

tively,

$$A_N^{(1)}(\beta,\theta) = \frac{1}{\sqrt{N}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta) \left\{\tau - \mathbf{1}^-(e_{ij})\right\} x_{ij},$$

$$A_N^{(2)}(\beta,\theta) = \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta) \left\{\frac{q}{2\theta} - \frac{1}{2} u_i^T u_i\right\}.$$

Moreover, the remainder term is simply given by

$$\begin{aligned} R_N^{(1)}(\delta,\gamma|y) = \frac{\delta^T}{\sqrt{N}} \int_0^1 \mathrm{d}t\, \nabla_\beta \ell\big(\beta + tN^{-\frac{1}{2}}\delta, \theta + tM^{-\frac{1}{2}}\gamma|y\big) - \nabla_\beta \ell(\beta,\theta|y) \\ + \frac{\gamma}{\sqrt{M}} \int_0^1 \mathrm{d}t\, \frac{\partial}{\partial\theta} \ell\big(\beta + tN^{-\frac{1}{2}}\delta, \theta + tM^{-\frac{1}{2}}\gamma|y\big) - \frac{\partial}{\partial\theta}\ell(\beta,\theta|y). \end{aligned}$$

Due to our model assumptions from Section 5.2.1 and Section 5.2.2, it is immediate to check that

$$\mathsf{E}^{\beta,\theta}\big[\Xi^T A_N(\beta,\theta)\big] = 0.$$

Furthermore, we may compute

$$\begin{aligned} \mathbf{Var}^{\beta,\theta}\big[\Xi^T A_N(\beta,\theta)\big] = \mathbf{Var}^{\beta,\theta}\big[\delta^T A_N^{(1)}(\beta,\theta)\big] + \mathbf{Var}^{\beta,\theta}\big[\gamma A_N^{(2)}(\beta,\theta)\big] \\ + 2\mathbf{Cov}^{\beta,\theta}\big[\delta^T A_N^{(1)}(\beta,\theta) A_N^{(2)}(\beta,\theta)\gamma\big]. \end{aligned}$$

Recall now that it is also part of our model assumptions from Section 5.2.1 that the response variable is independent across the clusters, or in other words that dependencies only occur within the clusters. This implies that the first-order approximation $A_N(\beta,\theta)$ is a sum of M independent random variables. Thus, for the first variance term we obtain

$$\begin{aligned} &\mathbf{Var}^{\beta,\theta}\big[\delta^T A_N^{(1)}(\beta,\theta)\big] \\ &= \frac{1}{N} \sum_{i=1}^{M} \sum_{j,k=1}^{n_i} \Big\{\mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ij})\big|Y_i=\cdot\big] \times \mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ik})\big|Y_i=\cdot\big]\Big] - \tau^2\Big\} \delta^T x_{ij} x_{ik}^T \delta \end{aligned}$$

and the second term is calculated as follows:

$$\mathbf{Var}^{\beta,\theta}\big[\gamma A_N^{(2)}(\beta,\theta)\big] = \frac{1}{M} \sum_{i=1}^{M} \left\{\mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[U_i^T U_i|Y_i=\cdot\big]^2\Big] - \frac{q^2}{\theta^2}\right\} \frac{\gamma^2}{4}.$$

The last term is given by

$$\begin{aligned} &\mathbf{Cov}^{\beta,\theta}\big[\delta^T A_N^{(1)}(\beta,\theta) A_N^{(2)}(\beta,\theta)\gamma\big] \\ &= \frac{1}{\sqrt{N}\sqrt{M}} \sum_{i=1}^{M} \sum_{j=1}^{n_i} \Big\{\mathsf{E}^{\beta,\theta}\Big[\mathsf{E}\big[\mathbf{1}^-(\varepsilon_{ij})\big|Y_i=\cdot\big] \times \mathsf{E}\big[U_i^T U_i\big|Y_i=\cdot\big]\Big] - \frac{q}{\theta}\tau\Big\} \delta^T x_{ij} \frac{\gamma}{2}. \end{aligned}$$

Hence, by Assumptions 3–5 we deduce that $\mathbf{Var}^{\beta,\theta}\big[\Xi^T A_N(\beta,\theta)\big] \to \Xi^T \mathbf{B} \Xi$ where $\mathbf{B} \in \mathbf{R}^{p+1\times p+1}$ is a positive definite matrix.

In a next step, let us check that Lyapunov's condition is satisfied for the scheme of random variables constituting the $\Xi^T A_N$'s. To this end, it is sufficient to find suitable bounds for the third moments of the random variables $\Xi^T \widetilde{A}_i(y)$ defined by

$$\int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta)\left(\sum_{j=1}^{n_i}\big\{\tau-\mathbf{1}^-(e_{ij})\big\}\frac{\delta^T x_{ij}}{\sqrt{N}}+\frac{\gamma}{\sqrt{M}}\Big\{\frac{q}{2\theta}-\frac{1}{2}u_i^T u_i\Big\}\right).$$

But an application of Jensen's inequality immediately shows that

$$\mathsf{E}^{\beta,\theta}\big[|\Xi^T\widetilde{A}_i|^3\big] \lesssim \left(\sum_{j=1}^{n_i}\frac{\|x_{ij}\|}{\sqrt{N}}\right)^3+\frac{1}{\sqrt{M}}\left(\sum_{j=1}^{n_i}\frac{\|x_{ij}\|}{\sqrt{N}}\right)^2+\frac{1}{M}\sum_{j=1}^{n_i}\frac{\|x_{ij}\|}{\sqrt{N}}+\frac{1}{M^{\frac{3}{2}}}.$$

In particular, Assumption 2 then guarantees that

$$\sum_{i=1}^{M}\mathsf{E}^{\beta,\theta}\big[|\Xi^T\widetilde{A}_i|^3\big]\to 0$$

as $N\to\infty$. As we have already seen that the variance term converges, this is enough in order to satisfy Lyapunov's condition and we can finally deduce that

$$\mathsf{P}^{\beta,\theta}\circ A_N(\beta,\theta)^{-1}\to\mathcal{N}(0,\mathbf{B}).$$

Step 2 (Second-order approximation): In this step, we push our expansion further and begin by introducing averages

$$R_N^{(1)}(\delta,\gamma|y)=\mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big]+R_N^{(1)}(\delta,\gamma|y)-\mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big].$$

In order to obtain a suitable second-order approximation, this is necessary due to the low regularity of the objective function of quantile regression (cf. Section 5.3). In fact, the disintegration of the response then enables us to Taylor expand with the notation from (5.18) as follows:

$$\begin{aligned}&\mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big]\\&=-\frac{1}{2}\delta^T B_N^{(1)}(\beta,\theta)\delta-\frac{1}{2}\gamma B_N^{(2)}(\beta,\theta)\gamma-\delta^T B_N^{(3)}(\beta,\theta)\gamma+R_N^{(2)}(\delta,\gamma)\\&=-\frac{1}{2}\Xi^T\mathbf{B}\Xi-\frac{1}{2}\Xi^T(B_N(\beta,\theta)-\mathbf{B})\Xi+R_N^{(2)}(\delta,\gamma).\end{aligned}$$

The remainder term $R_N^{(2)}(\delta,\gamma)$ is given by

$$R_N^{(2)}(\delta,\gamma)=-\int_0^1\mathrm{d}t\,(1-t)\,\Xi^T\big\{B_N\big(\beta+tN^{-\frac{1}{2}}\delta,\theta+tM^{-\frac{1}{2}}\gamma\big)-B_N(\beta,\theta)\big\}\Xi.$$

Step 3 (Analysis of remainder terms): In this part, we want to check the uniformity condition (5.25). To this end, we derive bounds for the different remainder terms due to

i) the limit covariance matrix approximation $\frac{1}{2}\Xi^T(B_N(\beta,\theta)-\mathbf{B})\Xi$,

ii) the second-order Taylor expansion $R_N^{(2)}(\delta,\gamma)$, and

iii) the disintegration of the response $R_N^{(1)}(\delta,\gamma|y)-\mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big]$.

All of these terms together constitute what is denoted in (5.22) as $R_N^{(3)}(\delta,\gamma|y)$. Furthermore, in the course of deriving the bounds we will determine how to choose the sequence $(\kappa_N)_{N\in\mathbf{N}}$ subject to (5.23).

Case i) We begin our analysis with the limit covariance approximation as it is the easiest case. We know that by definition

$$\kappa_N^{(1)} := \|B_N(\beta,\theta)-\mathbf{B}\|^{-\frac{1}{4}} \to \infty \quad \text{as} \quad N\to\infty.$$

It then follows directly by construction that, as $N\to\infty$,

$$\sup_{\|\Xi\|\le\kappa_N^{(1)}} \big|\Xi^T(B_N(\beta,\theta)-\mathbf{B})\Xi\big| \le \sup_{\|\Xi\|\le\kappa_N^{(1)}} \|\Xi\|^2\|B_N(\beta,\theta)-\mathbf{B}\| \to 0. \tag{5.26}$$

Case ii) Next, we study the remainder terms which arise due to the second-order Taylor expansion. In order to streamline our notation we will write (with $t\in(0,1)$)

$$\delta_N=\frac{\delta}{\sqrt{N}}, \quad \beta(t,\delta_N)=\beta+t\delta_N, \quad \gamma_N=\frac{\gamma}{\sqrt{M}}, \quad \theta(t,\gamma_N)=\theta+t\gamma_N.$$

Note that as a consequence of (5.24)—in other words, that we may assume without loss of generality that N is already sufficiently large—we have the non-degeneracy condition

$$\inf_{|\gamma|\le\kappa_N} \theta(t,\gamma_N) \ge \theta_- := \frac{\theta}{2} > 0. \tag{5.27}$$

Recall now that

$$\begin{aligned} R_N^{(2)}(\delta,\gamma) &= -\int_0^1 \mathrm{d}t\,(1-t)\,\Xi^T\big\{B_N\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)-B_N(\beta,\theta)\big\}\Xi \\ &= -\int_0^1 \mathrm{d}t\,(1-t)\,\big\{\delta^T R_N^{(\beta,\beta)}(t)\delta+\delta^T R_N^{(\beta,\theta)}(t)\gamma+\gamma R_N^{(\theta,\theta)}(t)\gamma\big\}, \end{aligned}$$

where $R_N^{(\beta,\beta)}(t)$ is given by (recall the expression for the second derivative with respect to β from Lemma 5.3.2)

$$\sum_{i=1}^M \sum_{j,k=1}^{n_i} \bigg(\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta)\,\big\{R^{ij}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)\times R^{ik}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)\big\}$$

$$- \int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta) \left\{ R^{ij}(\beta,\theta) \times R^{ik}(\beta,\theta) \right\} \Big) \frac{x_{ij} x_{ik}^T}{N},$$

the quantity $R_N^{(\beta,\theta)}(t)$ is defined as (recall the expression from the mixed second derivative with respect to β and θ from Lemma 5.3.4)

$$\sum_{i=1}^{M} \sum_{j=1}^{n_i} \Big(\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta) \left\{ R^{ij}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) \times R^{i}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) \right\}$$
$$- \int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta) \left\{ R^{ij}(\beta,\theta) \times R^{i}(\beta,\theta) \right\} \Big) \frac{x_{ij}}{\sqrt{M}\sqrt{N}}$$

and $R_N^{(\theta,\theta)}(t)$ is given by (recall the expression from the second derivative with respect to θ from Lemma 5.3.3)

$$\sum_{i=1}^{M} \int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta) \left\{ R^{i}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)^2 - R^{i}(\beta,\theta)^2 \right\} \frac{\gamma^2}{M}.$$

Here, we also made use of the shorthands

$$R^{ij}(\beta,\theta)(y) = \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta) \left\{ \tau - \mathbf{1}^-\big(e_{ij}(\beta)\big) \right\},$$
$$R^{i}(\beta,\theta)(y) = \int_{\mathbf{R}^q} \mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta) \, \frac{1}{2} \Big\{ \frac{q}{\theta} - \|u_i\|^2 \Big\}.$$

We proceed by estimating each of the three contributions separately. For the term $R_N^{(\theta,\theta)}(t)$, we may estimate due to Assumption 6 together with Hölder's inequality and Jensen's inequality

$$\big| R_N^{(\theta,\theta)}(t) \big| \lesssim \sum_{i=1}^{M} \int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta) \left| R^{i}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) - R^{i}(\beta,\theta) \right| \frac{1}{2M}.$$

In a next step, we decompose

$$\begin{aligned}
& f^{U_i|Y_i=y_i}\big(u_i|\beta(t,\delta_N),\theta(t,\gamma_N)\big) \Big\{ \frac{q}{\theta(t,\gamma_N)} - \|u_i\|^2 \Big\} \\
& - f^{U_i|Y_i=y_i}(u_i|\beta,\theta) \Big\{ \frac{q}{\theta} - \|u_i\|^2 \Big\} \\
& = f^{U_i|Y_i=y_i}(u_i|\beta,\theta) \Big\{ \frac{q}{\theta(t,\gamma_N)} - \frac{q}{\theta} \Big\} \\
& \quad + \big(f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta(t,\gamma_N)\big) - f^{U_i|Y_i=y_i}(u_i|\beta,\theta) \big) \Big\{ \frac{q}{\theta(t,\gamma_N)} - \|u_i\|^2 \Big\} \\
& \quad + \big(f^{U_i|Y_i=y_i}\big(u_i|\beta(t,\delta_N),\theta(t,\gamma_N)\big) - f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta(t,\gamma_N)\big) \big) \\
& \qquad\qquad \times \Big\{ \frac{q}{\theta(t,\gamma_N)} - \|u_i\|^2 \Big\}.
\end{aligned}$$

Now, let us estimate term by term. Recall from (5.24) the notation θ_-. With respect to the first

term, we may then derive based on the non-degeneracy condition (5.27) the bound

$$\begin{aligned}\Big|\int_{\mathbf{R}^q}\mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta)\left(\frac{q}{\theta(t,\gamma_N)}-\frac{q}{\theta}\right)\Big| &\lesssim \int_{\mathbf{R}^q}\mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\theta)\,\frac{t|\gamma_N|}{\theta_-^2}\\ &\lesssim |\gamma_N|.\end{aligned}$$

Furthermore, using Assumption 6 and again (5.27) we infer that the second term admits the upper bound

$$\begin{aligned}&\int_{\mathbf{R}^q}\mathrm{d}u_i\left|f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta(t,\gamma_N)\big)-f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta\big)\right|(1+\|u_i\|^2)\\ &\lesssim \sup_{|\widetilde{\theta}-\theta|\le 1\wedge\frac{\theta}{2}}\int_{\mathbf{R}^q}\mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\widetilde{\theta})\,(1+\|u_i\|^2)^2\frac{t|\gamma_N|}{\theta_-}\\ &\lesssim |\gamma_N|.\end{aligned}$$

Finally, due to the bound (5.11) and again Assumption 6 the third term is bounded from above by

$$\begin{aligned}&\int_{\mathbf{R}^q}\mathrm{d}u_i\left|f^{U_i|Y_i=y_i}\big(u_i|\beta(t,\delta_N),\theta(t,\gamma_N)\big)-f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta(t,\gamma_N)\big)\right|(1+\|u_i\|^2)\\ &\lesssim e^{C\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}}\int_{\mathbf{R}^q}\mathsf{P}^{U_i|Y_i=y_i}\big(\mathrm{d}u_i|\beta,\theta(t,\gamma_N)\big)\,(1+\|u_i\|^2)\sum_{j=1}^{n_i}\frac{t\|\delta\|\|x_{ij}\|}{\sqrt{N}}\\ &\quad+\sup_{(\widetilde{\beta},\widetilde{\theta})\in\mathfrak{K}(\beta,\theta)}\left(\int_{\mathbf{R}^q}\mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\widetilde{\beta},\widetilde{\theta})\,(1+\|u_i\|^2)\right)^2\sum_{j=1}^{n_i}\frac{t\|\delta\|\|x_{ij}\|}{\sqrt{N}}\\ &\lesssim e^{C\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}}\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}},\end{aligned}$$

where $C>0$ is a constant only depending on $\tau\in(0,1)$. All in all, these bounds lead to

$$\big|\gamma R_N^{(\theta,\theta)}(t)\gamma\big|\lesssim\frac{|\gamma|^3}{\sqrt{M}}+|\gamma|^2e^{C\max_i\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}}\max_{i=1,\ldots,M}\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}.$$

Thus, if we choose $\kappa_N^{(2)}:=M^{\frac{1}{24}}$ it then follows by Assumption 2

$$\kappa_N^{(2)}\to\infty\quad\text{and}\quad\sup_{\|\Xi\|\le\kappa_N^{(2)}}\big|\gamma R_N^{(\theta,\theta)}(t)\gamma\big|\to 0\quad\text{as}\quad N\to\infty. \tag{5.28}$$

We next turn to the term incorporating the second-order derivative with respect to the parameter β. Starting point for the corresponding term $R_N^{(\beta,\beta)}(t)$ is that it can be bounded from above by

$$\sum_{i=1}^{M}\sum_{j=1}^{n_i}\int_{\mathbf{R}^{n_i}}\mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta)\left|R^{ij}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)-R^{ij}(\beta,\theta)\right|\frac{\|x_{ij}\|}{\sqrt{N}}\sum_{k=1}^{n_i}\frac{\|x_{ik}\|}{\sqrt{N}}.$$

This time, we then make use of the decomposition

$$
\begin{aligned}
&f^{U_i|Y_i=y_i}\big(u_i|\beta(t,\delta_N),\theta(t,\gamma_N)\big)\big\{\tau-\mathbf{1}^-\big(e_{ij}(\beta+t\delta_N)\big)\big\}\\
&-f^{U_i|Y_i=y_i}(u_i|\beta,\theta)\big\{\tau-\mathbf{1}^-\big(e_{ij}(\beta)\big)\big\}\\
&=-f^{U_i|Y_i=y_i}(u_i|\beta,\theta)\big\{\mathbf{1}^-\big(e_{ij}(\beta+t\delta_N)\big)-\mathbf{1}^-\big(e_{ij}(\beta)\big)\big\}+\\
&\quad+\big(f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta(t,\gamma_N)\big)-f^{U_i|Y_i=y_i}(u_i|\beta,\theta)\big)\big\{\tau-\mathbf{1}^-\big(e_{ij}(\beta+t\delta_N)\big)\big\}\\
&\quad+\big(f^{U_i|Y_i=y_i}\big(u_i|\beta(t,\delta_N),\theta(t,\gamma_N)\big)-f^{U_i|Y_i=y_i}\big(u_i|\beta,\theta(t,\gamma_N)\big)\big)\\
&\qquad\qquad\qquad\qquad\times\big\{\tau-\mathbf{1}^-\big(e_{ij}(\beta+t\delta_N)\big)\big\}.
\end{aligned}
$$

After integration of the response, the first term appearing in this decomposition is bounded by

$$\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}$$

times an absolute constant. In addition, along the same lines as above we can bound the second term by

$$\sup_{|\widetilde{\theta}-\theta|\le 1\wedge\frac{\theta}{2}}\int_{\mathbf{R}^q}\mathsf{P}^{U_i|Y_i=y_i}(\mathrm{d}u_i|\beta,\widetilde{\theta})\,(1+\|u_i\|^2)\frac{t|\gamma_N|}{\theta_-}\lesssim|\gamma_N|,$$

and the last one by

$$e^{C\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}}\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}.$$

Hence, we obtain the bound

$$
\begin{aligned}
\big|\delta^T R_N^{(\beta,\beta)}(t)\delta\big|\lesssim\frac{\|\delta\|^3}{\sqrt{N}}e^{C\max_i\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}}\frac{1}{N}\sum_{i=1}^{M}\bigg(\sum_{j=1}^{n_i}\|x_{ij}\|\bigg)^3\\
+\frac{\|\delta\|^2|\gamma|}{\sqrt{N}}\max_{i=1,\ldots,M}\bigg(\sum_{j=1}^{n_i}\frac{\|x_{ij}\|}{N^{\frac{1}{4}}}\bigg)^2.
\end{aligned}
$$

Thus, if we choose $\kappa_N^{(3)}:=M^{\frac{1}{12}}$ we obtain by means of Assumption 2 that

$$\kappa_N^{(3)}\to\infty\quad\text{and}\quad\sup_{\|\Xi\|\le\kappa_N^{(3)}}\big|\delta^T R_N^{(\beta,\beta)}(t)\delta\big|\to 0\quad\text{as}\quad N\to\infty.\tag{5.29}$$

It remains to investigate the term $R_N^{(\beta,\theta)}(t)$. To this end, note that we can start estimating by means of Assumption 6 as follows:

$$\big|R_N^{(\beta,\theta)}(t)\big|$$

$$\lesssim \sum_{i=1}^{M}\sum_{j=1}^{n_i}\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta)\,\big|R^i\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)-R^i(\beta,\theta)\big|\frac{\|x_{ij}\|}{\sqrt{M}\sqrt{N}}$$
$$+\sum_{i=1}^{M}\sum_{j=1}^{n_i}\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta)\,\big|R^{ij}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big)-R^{ij}(\beta,\theta)\big|\frac{\|x_{ij}\|}{\sqrt{M}\sqrt{N}}.$$

Then, our arguments from above directly entail that

$$\big|\delta^T R_N^{(\beta,\theta)}(t)\gamma\big| \lesssim \frac{\|\delta\|^2|\gamma|}{\sqrt{M}\sqrt{N}} e^{C\max_i\sum_{j=1}^{n_i}\frac{\|\delta\|\|x_{ij}\|}{\sqrt{N}}} \max_{i=1,\dots,M}\left(\sum_{j=1}^{n_i}\frac{\|x_{ij}\|}{N^{\frac{1}{4}}}\right)^2$$
$$+\frac{\|\delta\||\gamma|^2}{\sqrt{M}N^{\frac{1}{4}}}\max_{i=1,\dots,M}\sum_{j=1}^{n_i}\frac{\|x_{ij}\|}{N^{\frac{1}{4}}}.$$

This motivates to define $\kappa_N^{(4)} := \kappa_N^{(3)}$ because then it holds

$$\kappa_N^{(4)}\to\infty \quad\text{and}\quad \sup_{\|\Xi\|\le\kappa_N^{(4)}}\big|\delta^T R_N^{(\beta,\theta)}(t)\gamma\big|\to 0 \quad\text{as}\quad N\to\infty \tag{5.30}$$

as a consequence of Assumption 2.

Now, simply put $\kappa_N^{(5)} = \kappa_N^{(2)}\wedge\kappa_N^{(3)}\wedge\kappa_N^{(4)}$ and observe that the convergence stated in (5.28), (5.29) and (5.30) is uniform with respect to $t\in[0,1]$. Hence,

$$\sup_{\|\Xi\|\le\kappa_N^{(5)}}\big|R_N^{(2)}(\delta,\gamma)\big|\to 0 \quad\text{as}\quad N\to\infty. \tag{5.31}$$

This concludes the discussion of case ii), i.e., the remainder term due to the second-order Taylor approximation.

Case iii) We move on with the analysis of the remainder term which originates from the disintegration of the response, i.e., we study the quantity

$$R_N^{(1)}(\delta,\gamma|y) - \mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big].$$

Note that this term actually is a sum of centered independent random variables due to the independence of the response across clusters. Thus, it suffices to derive a bound for the variance.

For this, we compute by Jensen's inequality that

$$\mathbf{Var}^{\beta,\theta}\Big[R_N^{(1)}(\delta,\gamma|\cdot)-\mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big]\Big]\lesssim\sum_{i=1}^{M}\int_0^1\mathrm{d}t\,\mathsf{E}^{\beta,\theta}\Big[\big(Z_N^{i1}(\delta,\gamma)+Z_N^{i2}(\delta,\gamma)\big)^2\Big],$$

where

$$Z_N^{i1}(\delta,\gamma)(t,y) = \frac{1}{\sqrt{N}}\sum_{j=1}^{n_i}\big\{R^{ij}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) - R^{ij}(\beta,\theta)\big\}\delta^T x_{ij},$$
$$Z_N^{i2}(\delta,\gamma)(t,y) = \frac{\gamma}{\sqrt{M}}\big\{R^{i}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) - R^{i}(\beta,\theta)\big\}.$$

In particular, by exploiting Assumption 6 and Young's inequality we obtain the bound

$$\begin{aligned}
&\mathbf{Var}^{\beta,\theta}\Big[R_N^{(1)}(\delta,\gamma|\cdot) - \mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big]\Big] \\
&\lesssim \int_0^1 \mathrm{d}t \sum_{i=1}^{M}\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta)\,\big|R^{i}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) - R^{i}(\beta,\theta)\big|\frac{\gamma^2}{M} \\
&\quad + \int_0^1 \mathrm{d}t \sum_{i=1}^{M}\sum_{j=1}^{n_i}\int_{\mathbf{R}^{n_i}} \mathsf{P}^{Y_i}(\mathrm{d}y_i|\beta,\theta)\,\big|R^{ij}\big(\beta(t,\delta_N),\theta(t,\gamma_N)\big) - R^{ij}(\beta,\theta)\big| \\
&\qquad\qquad \times \frac{\|x_{ij}\|}{\sqrt{N}}\sum_{k=1}^{n_i}\frac{\|\delta\|^2\|x_{ik}\|}{\sqrt{N}}.
\end{aligned}$$

Thus, the arguments from case ii) show that if we set $\kappa_N^{(6)} = \kappa_N^{(2)} \wedge \kappa_N^{(3)} \wedge \kappa_N^{(4)}$ then we obtain, as $N \to \infty$,

$$\sup_{\|\Xi\|\leq\kappa_N^{(6)}} R_N^{(1)}(\delta,\gamma|\cdot) - \mathsf{E}^{\beta,\theta}\big[R_N^{(1)}(\delta,\gamma|\cdot)\big] \to 0 \quad \text{in probability under } \mathsf{P}^{\beta,\theta}. \tag{5.32}$$

This eventually concludes our analysis of the remainder terms.

Step 4 (Conclusion): We finally define

$$\kappa_N := \kappa_N^{(1)} \wedge \kappa_N^{(5)} \wedge \kappa_N^{(6)}.$$

Our discussion from the third step in combination with Assumption 2 shows that the sequence $(\kappa_N)_{N\in\mathbf{N}}$ satisfies all requirements listed in (5.23) and that

$$\sup_{\|\Xi\|\leq\kappa_N} \big|R_N^{(3)}(\delta,\gamma|\cdot)\big| \to 0 \tag{5.33}$$

holds true in probability under $\mathsf{P}^{\beta,\theta}$ as $N \to \infty$. In particular, the uniformity condition (5.25) is fulfilled. This in turn concludes the proof of Theorem 5.2.1 by an application of Lemma 5.4.1. □

5.5 Simulation study

The proof conducted above does not only show the asymptotic behavior of the maximum likelihood estimation of mixed quantile regression, but additionally yields a plug-in variance estimator (c.f. Appendix 5.7.1). In the following simulation study, we will compare the plug-in

variance estimator to an empirical estimate. For this purpose, we have defined the following data generating process:

$$y_{ij} = 100 + 2x_{1,ij} + x_{2,ij} + u_i + \varepsilon_{ij}$$

with $i = 1, \ldots, M$, $j = 1, \ldots, 5$ and the error terms independently drawn from similar centered normal distributions with variances of 5. The explanatory variables consist of a metric and a dummy variable: $x_{1,ij} \sim U(0,1)$; $x_{2,ij} \sim Ber(0.5)$. While the cluster sizes are constant, we conduct simulations for $M = 5,\ 25,\ 50,\ 100,$ and 200, thus resulting in sample sizes of $N = 25,\ 125,\ 250,\ 500,$ and 1000. We generate 2000 independent samples under each setting. In every sample, we estimate the quantile regression models for $\tau \in \{0.05, 0.10, 0.25, 0.50\}$ and calculate the empirical variance of each parameter over the 2000 iterations. The resulting variances, respectively, their square roots serve as reference values for the plug-in standard errors. For the standard errors, we randomly draw 100 regression models out of the each iteration and calculate the respective standard errors as described in Appendix 5.7.1.

Table 5.1: Mean of plug-in estimates for the standard errors and their corresponding Monte Carlo reference values

	M	5		25		50		100		200	
τ	Covariate	Ref.	Plug.	Ref.	Plug.	Ref.	Plug.	Ref.	Plug.	Ref.	Plug.
0.05	Intercept	5.42	4.68	1.37	0.88	0.91	0.61	0.60	0.42	0.43	0.30
	Metric	3.52	6.81	1.30	1.41	0.95	1.03	0.61	0.70	0.45	0.48
	Dummy	2.36	5.08	0.78	0.79	0.54	0.57	0.38	0.39	0.26	0.28
0.10	Intercept	3.58	2.34	0.85	0.70	0.60	0.51	0.44	0.35	0.33	0.25
	Metric	2.61	3.45	1.04	1.12	0.77	0.87	0.51	0.59	0.38	0.40
	Dummy	1.79	2.44	0.63	0.63	0.45	0.48	0.32	0.33	0.23	0.23
0.25	Intercept	2.12	1.35	0.73	0.57	0.51	0.36	0.37	0.26	0.27	0.19
	Metric	1.96	2.06	0.89	0.90	0.67	0.63	0.45	0.44	0.32	0.31
	Dummy	1.43	1.16	0.52	0.50	0.37	0.36	0.26	0.24	0.19	0.18
0.50	Intercept	1.81	1.61	0.70	0.82	0.49	0.34	0.35	0.26	0.25	0.19
	Metric	1.85	2.64	0.84	1.17	0.62	0.61	0.42	0.45	0.30	0.32
	Dummy	1.36	0.75	0.49	0.59	0.34	0.38	0.24	0.23	0.17	0.18

In Table 5.1, we show the means of the resulting plug-in estimates with their respective empirical counterparts. In the settings with $M = 5$, the plug-in estimator seems to be very unreliable. Nevertheless, we see a strong improvement with increased sample sizes over all quantiles and explanatory variables, e.g., in the dummy variable, the standard errors differs by 0.65 at $M = 5$ and $\tau = 0.10$, but are virtually equal at $M = 200$. As expected, we observe estimates of quantiles closer to the center of the distribution to be more reliable, as their standard errors are substantially smaller. In respect of the covariates, the estimate for the dummy and the metric behave similarly well, while the standard error estimates for the intercept appear to be less

reliable. However, thinking about statistical inference, the standard error of the intercept is certainly only of subordinate importance.

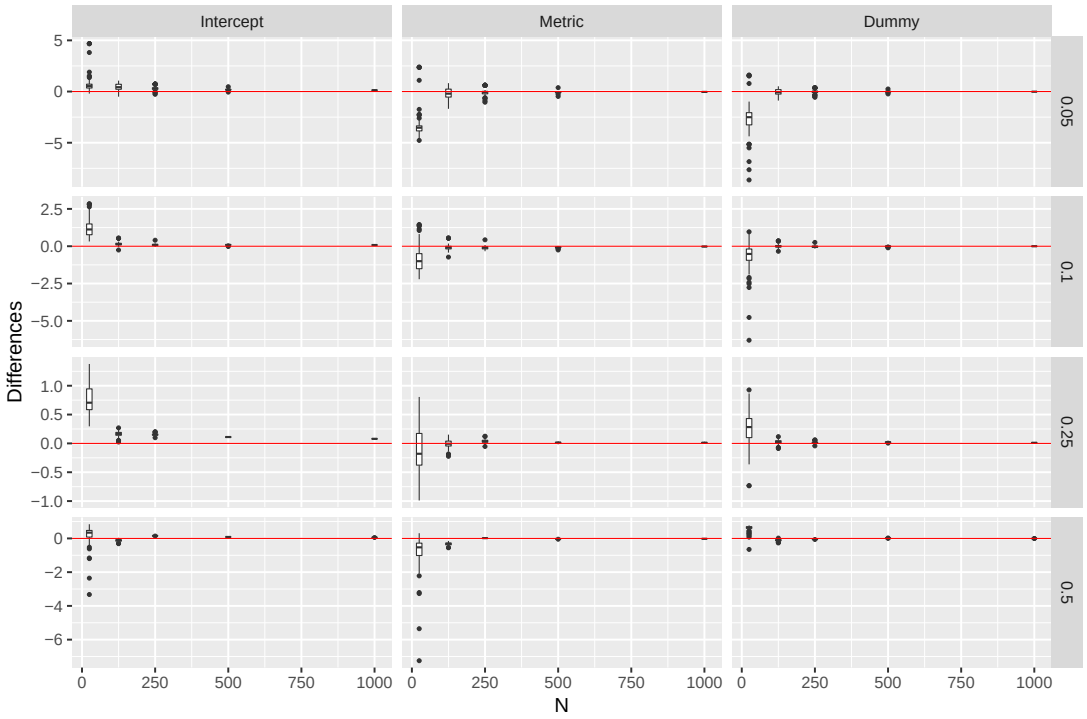


Figure 5.1: Boxplots of the differences between the plug-in standard error estimates and their corresponding Monte Carlo references. The red lines indicate the 0

Figure 5.1 shows the differences between plug-in standard errors and their corresponding empirical values in form of boxplots, grouped by explanatory variable, and the quantile. For all variables, it becomes clearly visible that the differences decrease towards zero with increased sample size. Furthermore, the spread of the boxplots decreases drastically, highlighting an increased precision in the plug-in estimator.

5.6 Conclusions and future research

We have shown the maximum likelihood estimator for the linear quantile mixed model to be consistent and asymptotically normally distributed. Additionally, we have derived a plug-in estimator of the corresponding covariance matrix. We have also demonstrated the plug-in estimator is accurate, except for very small samples. Therefore, we provide the results needed to apply standard tools of statistical inference, such as testing or confidence intervals.

In the future, an extension of the asymptotic theory to nonlinear quantile mixed models (Geraci, 2017) would be of highly interesting. Furthermore, extending count data quantile models (Machado and Silva, 2005) to the mixed case would certainly be of great benefit.

5.7 Appendix

5.7.1 Plug-in estimator for asymptotic covariance matrix

Let $\beta \in \mathbf{R}^p$ and $\sigma > 0$ be the unknown parameters, let M be the number of clusters and $n_1, \ldots, n_M$ the within-cluster sample sizes. In particular, $N = \sum_{i=1}^{M} n_i$ denotes the overall sample size. The sample observations are decomposed according to this structure, i.e. we write $y = y, \ldots, y_M \in \mathbf{R}^N$ with

$$y_i = (y_{i1}, \ldots, y_{in_i}) \in \mathbf{R}^{n_i}, \quad i = 1, \ldots M.$$

Analogously, we write $x_{ij} \in \mathbf{R}^p$ for the sample observations of the covariates with respect to the jth individual in the ith cluster. Finally, we write $\mathbf{1}^-$ for the characteristic function of the set $(-\infty, 0)$ and ρ_τ for the objective function of quantile regression. Recall that the model reads as follows

$$Y_{ij} = x_{ij}^T \beta_\tau + z_{ij}^T U_i + \varepsilon_{ij}$$

and that Y_i has conditional density given the value $U_i = u_i$

$$f^{Y_i|U_i=u_i} = \tau^{n_i}(1-\tau)^{n_i} \exp\left(-\sum_{j=1}^{n_i} \rho_\tau(y_{ij} - x_{ij}^T \beta_\tau + z_{ij}^T u_i) \right).$$

In particular, Y_i has density

$$f^{Y_i}(y_i) = \int_{\mathbf{R}^q} \mathrm{d}u_i f^{Y_i|U_i=u_i}(y_i) f^{U_i}(u_i)$$

where the density of the q-dimensional random effects is given by a q-dimensional Gaussian with covariance structure σ^2Id.

Now, we have everything in place in order to express the plug-in estimator for the asymptotic covariance matrix in terms of integrals. First, we obtain

$$\begin{aligned} \mathsf{E}\left[\mathbf{1}^-(\varepsilon_{ij}|Y_i = y_i)\right] &= \int_{\mathbf{R}^q} \mathrm{d}u_i f^{U_i|Y_i=y_i}(u_i) \mathbf{1}^-\left(x_{ij}^T \beta_\tau + z_{ij}^T u_i\right) \\ &= \int_{\mathbf{R}^q} \mathrm{d}u_i \frac{f^{Y_i|U_i=u_i}(y_i) f^{U_i}(u_i)}{f^{Y_i}(y_i)} \mathbf{1}^-\left(x_{ij}^T \beta_\tau + z_{ij}^T u_i\right) \end{aligned}$$

and similarly

$$\mathsf{E}\left[U_i^T U_i | Y_i = y_i\right] = \int_{\mathbf{R}^q} \mathrm{d}u_i \frac{f^{Y_i|U_i=u_i}(y_i) f^{U_i}(u_i)}{f^{Y_i}(y_i)} u_i^T u_i.$$

From this we deduce the expressions

$$\mathsf{E}^{\beta,\sigma}\Big[\mathsf{E}\left[\mathbf{1}^-(\varepsilon_{ij})|Y_i = \cdot\right] \times \mathsf{E}\left[\mathbf{1}^-(\varepsilon_{ik})|Y_i = \cdot\right]\Big]$$

$$= \int_{\mathbf{R}^{n_i}} \mathrm{d}y_i f^{Y_i}(y_i) \mathsf{E}\left[\mathbf{1}^-(\varepsilon_{ij})|Y_i = y_i\right] \times \mathsf{E}\left[\mathbf{1}^-(\varepsilon_{ik})|Y_i = y_i\right]$$

and

$$\mathsf{E}^{\beta,\sigma}\left[\mathsf{E}\left[U_i^T U_i|Y_i = \cdot\right]^2\right] = \int_{\mathbf{R}^{n_i}} \mathrm{d}y_i f^{Y_i}(y_i) \mathsf{E}\left[U_i^T U_i|Y_i = y_i\right]^2$$

For the plug-in estimator, just substitute the respective estimates for each occurrence of the true parameters.

Bibliography

Alfons, A. and M. Templ (2013). "Estimation of Social Exclusion Indicators from Complex Surveys: The R Package **laeken**". In: *Journal of Statistical Software* 54.15, pp. 1–25.

Alfons, A., M. Templ, and P. Filzmoser (2010). "An Object-Oriented Framework for Statistical Simulation: The R package **simFrame**". In: *Journal of Statistical Software* 37.3, pp. 1–36.

Basic, E. and U. Rendtel (2007). "Assessing the Bias Due to Non-Coverage of Residential Movers in the German Microcensus Panel: An Evaluation Using Data from the Socio-Economic Panel". In: *AStA Advances in Statistical Analysis* 91.3, pp. 311–334.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data". In: *Journal of the American Statistical Association* 83.401, pp. 28–36.

Bean, J. P. (1982). "Student Attrition, Intentions, and Confidence: Interaction Effects in a Path Model." In: *Research in Higher Education* 17.4, pp. 291–320.

Bedoya, H., S. Freije, L. Vila, G. Echeverria, D. Biller, G. M. Grandolini, R. Albisetti, E. Quintrell, and R. Vish (2013). *Country Partnership Strategy for the United Mexican States (2014-2019)*. Tech. rep. World Bank Group, Washington DC.

Berens, J., K. Schneider, S. Gortz, S. Oster, and J. Burghoff (2019). "Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods". In: *Journal of Educational Data Mining* 11.3, pp. 1–41.

Bertarelli, G., F. Schirripa Spagnolo, N. Salvati, and M. Pratesi (2019). "Small Area Estimation of Agricultural Data". In: *Spatial Econometric Methods in Agricultural Economics Using R*. accepted to be published. CRC book.

Bhageshpur, K. (2019). *Data Is The New Oil – And That's A Good Thing*. [accessed: 31.10.2020]. Forbes Media LLC. URL: https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/?sh=3508d2127304#4aa5bebd7304.

Bickel, P. J. and K. A. Doksum (1981). "An Analysis of Transformations Revisited". In: *Journal of the American Statistical Association* 76 (374), pp. 296–311.

Bivand, R. S. and D. W. S. Wong (2018). "Comparing Implementations of Global and Local Indicators of Spatial Association". In: *TEST* 27.3, pp. 716–748.

Boonstra, H. (2012). ***hbsae**: Hierarchical Bayesian Small Area Estimation*. R package version 1.0. R package version 1.0.

Box, G. and D. Cox (1964). "An Analysis of Transformations". In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 26.2, pp. 211–252.

Brandstätter, H., L. Grillich, and A. Farthofer (2006). "Prognose des Studienabbruchs". In: *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie* 38.3, pp. 121–131.

Breidenbach, J. (2015). *JoSAE: Functions for Some Unit-Level Small Area Estimators and their Variances*. R package version 0.2.3.

Breslow, N. E. and D. G. Clayton (1993). "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association* 88.421, pp. 9–25.

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". In: *Proceedings of Statistics Canada Symposium*.

Bundesamt für Eich- und Vermessungswesen (2017). *Verwaltungsgrenzen (VGD) - 1:250.000 Bezirksgrenzen, Daten vom 01.04.2017 von SynerGIS*. [accessed: 07.02.2018]. URL: http://data-synergis.opendata.arcgis.com/datasets/bb4acc011100469185d2e59fa4cae5fc_0.

Carroll, R. J. and D. Ruppert (1987). "Diagnostics and Robust Estimation when Transforming the Regression Model and the Response". In: *Technometrics* 29 (3), pp. 287–299.

Casas-Cordero, C., J. Encina, and P. Lahiri (2016). "Poverty Mapping for the Chilean Comunas". In: Pratesi, M. *Analysis of Poverty by Small Area Estimation*. Ed. by M. Pratesi. Wiley, pp. 379–403.

Chakravarti, I. M. and R. G. Laha (1967). "Handbook of Methods of Applied Statistics". In: *Handbook of methods of applied statistics*. John Wiley & Sons.

Chambers, J. and T. Hastie, eds. (1992). *Statistical Models in S*. Chapman & Hall, London.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). "Outlier Robust Small Area Estimation". In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 76.1, pp. 47–69.

Chambers, R. and N. Tzavidis (2006). "M-Quantile Models for Small Area Estimation". In: *Biometrika* 93.2, pp. 255–268.

Chandra, H., N. Salvati, and R. Chambers (2015). "A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation". In: *Journal of the Survey Statistics and Methodology* 3.2, pp. 109–135.

Chen, S. and P. Lahiri (2002). "A Weighted Jackknife MSPE Estimator in Small-Area Estimation". In: *Proceeding of the Section on Survey Research Methods*. American Statistical Association, pp. 473–477.

Cliff, A. and J. Ord (1981). *Spatial Processes: Models and Applications*. Pion, London.

CONEVAL (2010). *Methodology for Multidimensional Poverty Measurement in Mexico*. Report.

Cramér, H. (1928). "On the Composition of Elementary Errors". In: *Scandinavian Actuarial Journal* 1928.1, pp. 13–74.

Danilowicz-Gösele, K., K. Lerche, J. Meya, and R. Schwager (2017). "Determinants of Students' Success at University". In: *Education Economics* 25.5, pp. 513–532.

Das, K., M. Krzywinski, and N. Altman (2019). "Quantile Regression". In: *Nature Methods* 16, pp. 451–452.

Datta, G. S., R. E. Fay, and M. Ghosh (1991). "Hierarchical and Empirical Bayes Multivariate Analysis in Small Area Estimation". In: *Proceedings of Bureau of the Census 1991 Annual Research Conference*. Washington, DC: US Bureau of the Census, pp. 63–79.

Datta, G. S. and P. Lahiri (2000). "A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems". In: *Statistica Sinica* 10.2, pp. 613–627.

Datta, G., M. Ghosh, R. Steorts, and J. Maples (2011). "Bayesian Benchmarking with Applications to Small Area Estimation". In: *TEST* 20.3, pp. 574–588.

Diallo, M. and J. Rao (2014). *Small Area Estimation of Complex Parameters Under Unit-level Models with Skew-Normal Errors*. JSM 2014. Survey Research Methods Section.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). "Micro-Level Estimation of Poverty and Inequality". In: *Econometrica* 71.1, pp. 355–364.

Elbers, C. and R. van der Weide (2014). *Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality*. Paper is funded by the Knowledge for Change Program (KCP).

Eurostat (2004). *Common Cross-Sectional EU indicators Based on EU-SILC; the Gender Pay Gap*. EU-SILC 131-rev/04. Luxembourg: Unit D-2: Living conditions, social protection, Directorate D: Single Market, Employment, and Social statistics, Eurostat.

Fabrizi, E. and C. Trivisano (2016). "Small Area Estimation of the Gini Concentration Coefficient". In: *Computational Statistics and Data Analysis* 99, pp. 223–234.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer Science & Business Media.

Fay, R. E. and R. A. Herriot (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data". In: *Journal of the American Statistical Association* 74.366, pp. 269–277.

Feng, X., X. He, and J. Hu (2011). "Wild Bootstrap for Quantile Regression". In: *Biometrika* 98 (4), pp. 995–999.

Fleischer, J., D. Leutner, M. Brand, H. Fischer, M. Lang, P. Schmiemann, and E. Sumfleth (2019). "Vorhersage des Studienabbruchs in naturwissenschaftlich-technischen Studiengängen". In: *Zeitschrift für Erziehungswissenschaft* 22.5, pp. 1077–1097.

Foster, J., J. Greer, and E. Thorbecke (1984). "A Class of Decomposable Poverty Measures". In: *Econometrica* 52.3, pp. 761–766.

Georg, W. (2008). "Individuelle und institutionelle Faktoren der Bereitschaft zum Studienabbruch: eine Mehrebenenanalyse mit Daten des Konstanzer Studierendensurveys." In: *Zeitschrift für Soziologie der Erziehung und Sozialisation* 28.2, pp. 191–206.

Geraci, M. (2014). "Linear Quantile Mixed Models: The lqmm Package for Laplace Quantile Regression". In: *Journal of Statistical Software* 57.13, pp. 1–29.

Geraci, M. (2017). “Nonlinear quantile mixed models”. In: *arXiv preprint arXiv:1712.09981*.

Geraci, M. and M. Bottai (2007). “Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution”. In: *Biostatistics* 8.1, pp. 140–154.

— (2014). “Linear Quantile Mixed Models”. In: *Statistics and Computing* 24.3, pp. 461–479.

Geyer, C. F. R. (1996). “On the Asymptotics of Convex Stochastic Optimization”. In.

Gold, A. (1988). *Studienabbruch, Abbruchneigung und Studienerfolg. Vergleichende Bedingungsanalysen des Studienverlaufs*. Europäische Hochschulschriften.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2008). “Bootstrap Mean Squared Error of a Small-Area EBLUP”. In: *Journal of Statistical Computation and Simulation* 78 (5), pp. 443–462.

Graf, M., J. M. Marín, and I. Molina (2019). “A Generalized Mixed Model for Skewed Distributions Applied to Small Area Estimation”. In: *TEST* 28.2, pp. 565–597.

Gurka, M. J., L. . Edwards, K. Muller, and L. L. Kupper (2006). “Extending the Box-Cox Transformation to the Linear Mixed Model”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.2, pp. 273–288.

Hadam, S., N. Würz, and A.-K. Kreutzmann (2020). “Estimating Regional Unemployment with Mobile Network Data for Functional Urban Areas in Germany”. In: *Refubium - Freie Universität Berlin Repository*, pp. 1–28.

Hagenaars, A., K. de Vos, and M. Zaidi (1994). *Poverty Statistics in the Late 1980s: Research Based on Mirco-data*. Office for the Official Publications of the European Communities.

Hahm, S. and J. Storck (2018). “Das Potenzial administrativer Daten für das Qualitätsmanagement an Hochschulen”. In: *Zeitschrift für Hochschulentwicklung* 13.1, pp. 193–207.

Heublein, U., J. Ebert, C. Hutzsch, S. Isleib, R. König, J. Richter, and A. Woisch (2017). *Zwischen Studienerwartungen und Studienwirklichkeit: Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen*. Deutsches Zentrum für Hochschul-und Wissenschaftsforschung (DZHW).

Himmler, O., R. Jäckle, and P. Weinschenk (2019). “Soft Commitments, Reminders, and Academic Performance”. In: *American Economic Journal: Applied Economics* 11.2, pp. 114–42.

Hinneberg, H. (2003). “Abiturnote und Studienerfolg”. In: *Hochschulwesen* 51, pp. 145–146.

Hjort, N. L. and D. Pollard (1993). “Asymptotics for Minimisers of Convex Processes”. In: *arXiv preprint arXiv:1107.3806*.

Hoeting, J. A. and J. G. Ibrahim (1998). “Bayesian Predictive Simultaneous Variable and Transformation Selection in the Linear Model”. In: *Computational Statistics & Data Analysis* 28.1, pp. 87–103.

Hoeting, J. A., A. E. Raftery, and D. Madigan (2002). “Bayesian Variable and Transformation Selection in Linear Regression”. In: *Journal of Computational and Graphical Statistics* 3.3, pp. 485–507.

Horvitz, D. and D. Thompson (1952). "A Generalization of Sampling Without Replacement from a Finite Universe". In: *Journal of the American Statistical Association* 47.260, pp. 663–685.

Isphording, I. E. and F. Wozny (2018). *Ursachen des Studienabbruchs–eine Analyse des Nationalen Bildungspanels*. Tech. rep. Institute for the Study of Labor (IZA).

Jiang, J., P. Lahiri, and S.-M. Wan (2002). "A Unified Jackknife Theory for Empirical Best Prediction with M-Estimation". In: *The Annals of Statistics* 30.6, pp. 1782–1810.

Jiang, J., P. Lahiri, S.-M. Wan, and C.-H. Wu (2001). "Jackknifing in the Fay-Herriot Model with an Example". In: *Proceedings of the Seminar on Funding Opportunity in Survey Research Council of Professional Associations on Federal Statistics*. Washington DC: Bureau of Labor Statistics, pp. 75–79.

Jiang, J. and J. S. Rao (2020). "Robust Small Area Estimation: An Overview". In: *Annual Review of Statistics and Its Application* 7, pp. 337–360.

John, J. A. and N. R. Draper (1980). "An Alternative Family of Transformations". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 29 (2), pp. 190–197.

Jung, S.-H. (1996). "Quasi-Likelihood for Median Regression Models". In: *Journal of the American Statistical Association* 91.433, pp. 251–257.

Knight, K. (1998). "Limiting Distributions for L1 Regression Estimators under General Conditions". In: *Annals of Statistics*, pp. 755–770.

— (2003). "On the Second Order Behaviour of the Bootstrap L1 Regression Estimators". In: *Journal of The Iranian Statistical Society* 2.1, pp. 21–42.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. and G. Bassett Jr (1978). "Regression Quantiles". In: *Econometrica*, pp. 33–50.

Koenker, R. and J. A. F. Machado (1999). "Goodness of Fit and Related Inference Processes for Quantile Regression". In: *Journal of the American Statistical Association* 94.448, pp. 1296–1310.

Komsta, L. and F. Novomestky (2015). ***moments**: Moments, Cumulants, Skewness, Kurtosis and Related Tests*. R package version 0.14.

Kreutzmann, A.-K., P. Marek, N. Salvati, and T. Schmid (2019). "Estimating Regional Wealth in Germany: How Different are East and West Really?" In: *Bundesbank Discussion Paper* 35/2019.

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". In: *Journal of Statistical Software* 91.7, pp. 1–33.

Kühn, M., L. Uta, and P. Klaus (2018). *Erfolgsbias in Datenbeständen der empirischen Bildungsforschung?: Eine Analyse auf Basis der NEPS-Daten.* Tech. rep. 6. Tagung der Gesellschaft für Empirische Bildungsforschung.

Lahiri, P. and J. Suntornchost (2015). "Variable Selection for Linear Mixed Models with Applications in Small Area Estimation". In: *The Indian Journal of Statistics* 77-B.2, pp. 312–320.

Laud, P. W. and J. G. Ibrahim (1995). "Predictive Model Selection". In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 57.1, pp. 247–262.

Lefler, M., D. Gonzalez, and A. Martin (2014). ***saery**: Small Area Estimation for Rao and Yu Model*. R package version 1.0. R package version 1.0.

Li, H. and P. Lahiri (2010). "An Adjusted Maximum Likelihood Method for Solving Small Area Estimation Problems". In: *Journal of Multivariate Analyis* 101.4, pp. 882–902.

Lopez-Vizcaino, E., M. Lombardia, and D. Morales (2019). ***mme**: Multinomial Mixed Effects Models*. R package version 0.1-6. R package version 0.1-6.

López-Vizcaíno, E., M. J. Lombardía, and D. Morales (2013). "Multinomial-Based Small Area Estimation of Labour Force Indicators". In: *Statistical modelling* 13.2, pp. 153–178.

Machado, J. A. F. and J. S. Silva (2005). "Quantiles for Counts". In: *Journal of the American Statistical Association* 100.472, pp. 1226–1237.

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". In: *Journal of Official Statistics* 31.2, pp. 263–281.

Marhuenda, Y., I. Molina, and D. Morales (2013). "Small Area Estimation with Spatio-Temporal Fay-Herriot Models". In: *Computational Statistics and Data Analysis* 58, pp. 308–325.

Marhuenda, Y., D. Morales, and M. del Camen Pardo (2014). "Information Criteria for Fay-Herriot Model Selection". In: *Computational Statistics and Data Analysis* 70, pp. 268–280.

Marino, M. F., M. G. Ranalli, N. Salvati, and M. Alfo (2019). "Semi-Parametric Empirical Best Prediction for Small Area Estimation of Unemployment Indicators". In: *Annals of Applied Statistics*, forthcoming.

Marino, M. F., N. Tzavidis, and M. Alfo (2018). "Mixed hidden Markov Quantile Regression Models for Longitudinal Data with Possibly Incomplete Sequences". In: *Statistical Methods in Medical Research* 27.7, pp. 2231–2246.

McLachlan, G. J. and T. Krishnan (2007). *The EM Algorithm and Extensions*. Vol. 382. John Wiley & Sons.

Miller, J. J. (1977). "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance". In: *The Annals of Statistics* 5.4, pp. 746–762.

Miltiadou, M. (2020). *Measuring and Reporting Reliability of Labour Force Survey and Annual Population Survey Estimates Force Survey and Annual Population Survey Estimates*. [accessed: 05.06.2020]. UK Office for National Statistics. URL: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/measuringandreportingreliabilityoflabourforcesurveyandannualpopulationsurveyestimates.

Molina, I. and J. Rao (2010). "Small Area Estimation of Poverty Indicators". In: *The Canadian Journal of Statistics* 38.3, pp. 369–385.

Molina, I., J. Rao, and G. Datta (2015). "Small Area Estimation Under a Fay-Herriot Model with Preliminary Testing for the Presence of Random Area Effects". In: *Survey Methodology* 41.1, pp. 1–19.

Molina, I., N. Salvati, and M. Pratesi (2009). "Bootstrap for Estimating the MSE of the Spatial EBLUP". In: *Computational Statistics* 24, pp. 441–458.

Molina, I. and Y. Marhuenda (2015). "**sae**: An R Package for Small Area Estimation". In: *The R Journal* 7.1, pp. 81–98.

Molina, I. and N. Martín (2018). "Empirical Best Prediction Under a Nested Error Model with Log Transformation". In: *The Annals of Statistics* 46.5, pp. 1961–1993.

Molina, I., A. Saei, and M. José Lombardía (2007). "Small Area Estimates of Labour Force Participation Under a Multinomial Logit Mixed Model". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.4, pp. 975–1000.

Multrus, F., S. Majer, T. Bargel, and M. Schmidt (2017). *Studiensituation und studentische Orientierungen. 13. Studierendensurvey an Universitäten und Fachhochschulen.* Tech. rep. BMBF.

Nakagawa, S. and H. Schielzeth (2013). "A General and Simple Method for Obtaining R^2 from Generalized Linear Mixed-Effects Models". In: *Methods in Ecology and Evolution* 4, pp. 133–142.

Nandy, A. (2015). ***smallarea**: Fits a Fay-Herriot Model*. R package version 0.1. R package version 0.1.

Neugebauer, M., U. Heublein, and A. Daniel (Oct. 2019). "Studienabbruch in Deutschland: Ausmaß, Ursachen, Folgen, Präventionsmöglichkeiten". In: *Zeitschrift für Erziehungswissenschaft*. ISSN: 1862-5215.

Neves, A., D. Silva, and S. Correa (2013). "Small Domain Estimation for the Brazilian Service Sector Survey". In: *ESTADÍSTICA* 65.185, pp. 13–37.

Pebesma, E. (2018). "Simple Features for R: Standardized Support for Spatial Vector Data". In: *The R Journal* 10.1, pp. 439–446.

Pebesma, E. J. and R. S. Bivand (Nov. 2005). "Classes and Methods for Spatial Data in R". In: *R News* 5.2, pp. 9–13.

Permatasari, N. and A. Ubaidillah (2020). ***msae**: Multivariate Fay Herriot Models for Small Area Estimation*. R package version 0.1.1.

Petrucci, A. and N. Salvati (2006). "Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment". In: *Journal of Agricultural, Biological and Environmental Statistics* 11.2, pp. 169–182.

Pfeffermann, D. (2013). "New Important Developments in Small Area Estimation". In: *Statistical Science* 28.1, pp. 40–68.

Pinheiro, J. C. and E. C. Chao (2006). "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized linear mixed models". In: *Journal of Computational and Graphical Statistics* 15.1, pp. 58–81.

Pinheiro, J. C. (1994). "Topics in Mixed Effects Models". PhD thesis. University of Wisconsin, Madison.

Prasad, N. and J. Rao (1990). "The Estimation of the Mean Squared Error of Small-Area Estimation". In: *Journal of the American Statistical Association* 85.409, pp. 163–171.

Pratesi, M., ed. (2016). *Analysis of Poverty Data by Small Area Estimation*. Wiley.

Pratesi, M. and N. Salvati (2008). "Small Area Estimation: the EBLUP Estimator Based on Spatially Correlated Random Area Effects". In: *Statistical Methods and Applications* 17.1, pp. 113–141.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation*. New York: Wiley.

Rao, J. N. K. and M. Yu (1994). "Small-Area Estimation by Combining Time-Series and Cross-Sectional Data". In: *The Canadian Journal of Statistics* 22.4, pp. 511–528.

Rivest, L.-P. and N. Vandal (2003). "Mean Squared Error Estimation for Small Areas when the Small Area Variances are Estimated". In: *Proceedings of International Conference of Recent Advanced Survey Sampling*, pp. 197–206.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2020). "Data-Driven Transformations in Small Area Estimation". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 183.1, pp. 121–148.

Royston, P. and P. C. Lambert (2011). *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. StataCorp LP.

Saei, A. and R. Chambers (2005). "Out of Sample Estimation for Small Areas using Area Level Data". In: *Southampton Statistical Sciences Research Institute Methodology Working Paper* M05/11. Southampton Statistical Sciences Research Institute, UK.

Saei, A. and R. Chambers (2003). "Small Area Estimation Under Linear and Generalized Linear Mixed Models with Time and Area Effects". In: *Project Report M03/15*.

Sarcletti, A. and S. Müller (2011). "Zum Stand der Studienabbruchforschung. Theoretische Perspektiven, zentrale Ergebnisse und methodische Anforderungen an künftige Studien". In: *Zeitschrift für Bildungsforschung* 1.3, pp. 235–248.

Schall, R. (1991). "Estimation in Generalized Linear Models with Random Effects". In: *Biometrika* 78.4, pp. 719–727.

Schiefele, U. and L. Jacob-Ebbinghaus (2006). "Lernermerkmale und Lehrqualität als Bedingungen der Studienzufriedenheit". In: *Zeitschrift für Pädagogische Psychologie* 20.3, pp. 199–212.

Schimpl-Neimanns, B. (2008). *Bildungsverläufe und Stichprobenselektivität: Analysen zur Stichprobenselektivität des Mikrozensuspanels 1996-1999 am Beispiel bildungsstatistischer Fragestellungen*. Vol. 1. DEU.

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". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.4, pp. 1163–1190.

Schmid, T., N. Tzavidis, R. Münnich, and R. Chambers (2016). "Outlier Robust Small Area Estimation Under Spatial Correlation". In: *Scandinavian Journal of Statistics: Theory and Applications* 43.3, pp. 806–826.

Schneider, K., J. Berens, and J. Burghoff (2019). "Drohende Studienabbrüche durch Frühwarnsysteme erkennen: Welche Informationen sind relevant?" In: *Zeitschrift für Erziehungswissenschaft* 22.5, pp. 1121–1146.

Schröder-Gronostay, M. (1999). "Studienabbruch–Zusammenfassung des Forschungsstandes". In: *Studienerfolg und Studienabbruch*, pp. 209–240.

Shi, C. (2018). ***BayesSAE**: Bayesian Analysis of Small Area Estimation*. R package version 1.0-2. R package version 1.0-2.

Singh, B. B., K. Shukla, and D. Kundu (2005). "Spatio-Temporal Models in Small Area Estimation". In: *Survey Methodology* 31.2, pp. 183–195.

Sinha, S. and J. Rao (2009). "Robust Small Area Estimation". In: *The Canadian Journal of Statistics* 37.3, pp. 381–399.

Slud, E. and T. Maiti (2006). "Mean-Squared Error Estimation in Transformed Fay-Herriot Models". In: *Journal of the Royal Statistical Society Series B* 68.2, pp. 239–257.

Sugawasa, S. and T. Kubokawa (2017). "Transforming Response Values in Small Area Prediction". In: *Computational Statistics and Data Analysis* 114, pp. 47–60.

Tinto, V. (1975). "Dropout from Higher Education: A Theoretical Synthesis of Recent Research". In: *Review of educational research* 45.1, pp. 89–125.

Tortajada, C. (2006). *Who has Access to Water Case study of Mexico City Metropolitan Area Human Development Report 2006*. Tech. rep. United Nations Development Programme.

Trapmann, S., B. Hell, S. Weigand, and H. Schuler (2007). "Die Validität von Schulnoten zur Vorhersage des Studienerfolgs-eine Metaanalyse". In: *Zeitschrift für pädagogische Psychologie* 21.1, pp. 11–27.

Tzavidis, N., R. Chambers, N. Salvati, and H. Chandra (2012). "Small Area Estimation in Practice an Application to Agricultural Business Survey Data". In: *Journal of the Indian Society of Agricultural Statistics* 66.1, pp. 213–228.

Tzavidis, N., L.-C. Zhang, A. Luna Hernandez, T. Schmid, and N. Rojas-Perilla (2018). "From Start to Finish: A Framework for the Production of Small Area Official Statistics". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4, pp. 927–979.

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 0.5.0.

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Vol. 1. Springer Series in Statistics.

Walker, A. (2018). ***openxlsx**: Read, Write and Edit XLSX Files*. R package version 4.1.0.

Wang, J. and W. A. Fuller (2003). "The Mean Squared Error of Small Area Predictors Constructed With Estimated Area Variances". In: *Journal of the American Statistical Association* 98 (463), pp. 716–723.

Warnholz, S. (2016). “Small Area Estimation Using Robust Extensions to Area Level Models”. Freie Universität Berlin. PhD thesis. Freie Universität Berlin.

— (2018). ***saeRobust**: Robust Small Area Estimation*. R package version 0.2.0.

Weidenhammer, B., N. Tzavidis, T. Schmid, and N. Salvati (2014). “Domain Prediction for Counts Using Microsimulation via Quantiles”. In: *Small Area Estimation 2014 Conference*. Poznan, Poland.

Weidenhammer, B. (2016). “The Consistency of Quantile Regression in Linear Mixed Models”. eng. PhD thesis. Berlin: Freie Universität Berlin.

Weiss, L. (1971). “Asymptotic Properties of Maximum Likelihood Estimators in some Nonstandard Cases”. In: *Journal of the American Statistical Association* 66.334, pp. 345–350.

— (1973). “Asymptotic Properties of Maximum Likelihood Estimators in some Nonstandard Cases, II”. In: *Journal of the American Statistical Association* 68.342, pp. 428–430.

Wickham, H. et al. (2014). “Tidy Data”. In: *Journal of Statistical Software* 59.10, pp. 1–23.

Wickham, H. (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. Springer New York. ISBN: 978-3-319-24277-4.

Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani (2019). “Welcome to the tidyverse”. In: *Journal of Open Source Software* 4.43, p. 1686.

Wiers-Jenssen, J., B. Stensaker, and J. B. Grøgaard (2002). “Student Satisfaction: Towards an Empirical Deconstruction of the Concept”. In: *Quality in higher education* 8.2, pp. 183–195.

Yang, L. (1995). “Transformation-Density Estimation”. PhD thesis. University of North Carolina, Chapel Hill.

Yang, Z. (2006). “A Modified Family of Power Transformations”. In: *Economics Letters* 92, pp. 14–19.

Ybarra, L. M. R. and S. L. Lohr (2008). “Small Area Estimation When Auxiliary Information Is Measured with Error”. In: *Biometrika* 95.4, pp. 919–931.

Yeo, I.-K. and R. Johnson (2000). “A New Family of Power Transformations to Improve Normality or Symmetry”. In: *Biometrika* 87.4, pp. 954–959.

Yoshimori, M. and P. Lahiri (2014). “A New Adjusted Maximum Likelihood Method for the Fay-Herriot Small Area Model”. In: *Journal of Multivariate Analysis* 124, pp. 281–294.

You, Y. and B. Chapman (2006). “Small Area Estimation Using Area Level Models and Estimated Sampling Variances”. In: *Survey Methodology* 32.1, pp. 97–103.

Zhang, X., J. Holt, S. Yun, H. Lu, K. Greenlund, and J. Croft (2015). “Validation of Multilevel Regression and Poststratification Methodology for Small Area Estimation of Health Indicators From the Behavioral Risk Factor Surveillance System”. In: *American Journal of Epidemiology* 182.2 (2), pp. 127–137.

Summaries

Summaries in English

Abstract: Forecasting Study-Success or Drop-Out Based on Survey- and Administrative Examination Data: The Project "Students' Progressions" at the School of Business and Economics at FU Berlin

The measurement of study-success or drop-out is often done by retrospective analysis of data from students who already left the university. Such surveys, however, are subject to high non-response rates. The retrospective self-rating is also subject to measurement error. There are, alternatively, certain prospective approaches in the framework of panel surveys which are subject to losses between panel waves.

This article presents a new prospective approach on the basis of administrative examination data and survey data. Background information on students is surveyed in the classroom at the start of the second term. The student's mandatory consent for linking the surveyed information with the examination records is almost always achieved, such that it is possible to analyze the impacts of social background, part-time work during the studies and motivation for the subject on finalizing the study program can be analyzed without sample losses.

This approach was first carried out at the Economic Department of the FU Berlin. This article describes how this concept was realized and the empirical results achieved on the study program and students drop-out. From the collected credit points and the student's self-assessments, we can determine that it is possible to precisely predict whether the studies will be abandoned already at the initial stage of the study program. On the other hand, the school score and the individuals' social background do not deliver any additional information on a students drop-out.

Keywords: Students drop-out, prospective survey design, administrative examination data, longitudinal analysis

Abstract: Data-driven Transformations in Small Area Estimation

Small area models typically depend on the validity of model assumptions. For example, a commonly used version of the Empirical Best Predictor relies on the Gaussian assumptions of

the error terms of the linear mixed regression model, a feature rarely observed in applications with real data. The present paper proposes to tackle the potential lack of validity of the model assumptions by using data-driven scaled transformations as opposed to ad-hoc chosen transformations. Different types of transformations are explored, the estimation of the transformation parameters is studied in detail under the linear mixed regression model and transformations are used in small area prediction of linear and non-linear parameters. The use of scaled transformations is crucial, as it allows for fitting the linear mixed regression model with standard software and hence it simplifies the work of the data analyst. Mean squared error estimation that accounts for the uncertainty due to the estimation of the transformation parameters is explored using parametric and semi-parametric (wild) bootstrap. The proposed methods are illustrated using real survey and census data for estimating income deprivation parameters for municipalities in the Mexican state of Guerrero. Simulation studies and the results from the application show that using carefully selected, data-driven transformations can improve small area estimation.

Keywords: Random effects; bootstrap; adaptive transformations; maximum likelihood estimation; poverty mapping

Abstract: A Framework for Producing Small Area Estimates Based on Area-Level Models in R

The R package **emdi** facilitates the estimation of regionally disaggregated indicators using small area estimation methods and provides tools for model building, diagnostics, presenting, and exporting the results. The package version 1.1.7 includes unit-level small area models that rely on access to micro data which may be challenging due to confidentiality constraints. In contrast, area-level models are less demanding with respect to (a) data requirements, as only aggregates are needed for estimating regional indicators, and (b) computational resources, and enable the incorporation of design-based properties. Therefore, the area-level model (Fay and Herriot, 1979) and various extensions have been added to version 2.0.1 of the package **emdi**. These extensions include, among others, (a) transformed area-level models with back-transformations, (b) spatial and robust extensions, (c) adjusted variance estimation methods, and (d) area-level models that account for measurement errors. Corresponding mean squared error estimators are implemented for assessing the uncertainty. User-friendly tools, such as a stepwise variable selection function, model diagnostics, benchmarking options, high quality maps, and export options of the results enable the user a complete analysis procedure - from model building to diagnostics. The functionality of the package is demonstrated by illustrative examples based on synthetic data for districts in Austria.

Keywords: Fay-Herriot models, official statistics, survey statistics, small area estimation

Abstract: ammlogit: Estimation and Prediction Using an Aggregated Mixed Multinomial Logit Model

The R-package **ammlogit** implements a mixed multinomial logistic regression for aggregated data (Molina, Saei, et al., 2007). The original method has been extended to allow for out-of-sample domains as well as for varying information on the sample of survey and population. In addition to the model estimation, the user is assisted in the model assessment. R-typical methods are provided for the estimates. Ratios and absolute values per category are produced simultaneously as well on domain as on sub-group level by default. A parametric bootstrap has been implemented to provide mean squared error estimates for the predicted values. Furthermore, **ammlogit** provides assistance in the latter's visualization. Provided suitable shape files are available, **ammlogit** enables users to create high quality choropleth maps.

Keywords: R, mixed multinomial logistic regression, aggregated data, small area estimation, poverty mapping

Abstract: Asymptotic distribution of regression quantiles in a mixed effects model

Linear quantile models allow for a robust analysis of the conditional distribution of the variable of interest. The introduction of a random effects term extended their range of application to data with complex dependency structures, as they occur in many studies. This paper proposes a higher theoretical understanding of linear quantile mixed models by analyzing the asymptotic behavior of the corresponding maximum likelihood estimator. We will proof the estimators to be consistent and show that it is asymptotically normally distributed. Additionally, a plug-in variance estimator is derived and its finite sample behavior is demonstrated in a simulation study.

Keywords: Quantile regression, linear mixed effects models, asymptotic distribution, maximum likelihood, asymetric Laplace distribution

Zusammenfassungen auf Deutsch

Zusammenfassung: Die Prognose von Studienerfolg und Studienabbruch auf Basis von Umfrage- und administrativen Prüfungsdaten: Das Projekt „Studienverläufe“ am FB Wirtschaftswissenschaft der FU Berlin

Die Messung von Studienerfolg bzw. Studienabbruch erfolgt häufig retrospektiv anhand von Exmatrikulierten-Befragungen. Diese Erhebungen sind jedoch mit hohen Nonresponse-Raten verknüpft. Auch die retrospektive Selbsteinschätzung unterliegt Erinnerungsfehlern. Alternativ findet man auch prospektive Ansätze im Rahmen von Panelerhebungen, die jedoch von Stichprobenausfällen zwischen den Befragungswellen betroffen sind.

Dieser Artikel präsentiert einen neuen prospektiven Ansatz auf Basis von administrativen Prüfungsdaten und Umfragedaten. Hintergrundinformationen über die Studierenden werden zu Beginn des zweiten Fachsemesters im Rahmen einer Hörsaalbefragung erhoben. Die notwendige Einwilligung der Studierenden zur Verknüpfung mit den Prüfungsdaten wird fast immer erreicht, so dass der Einfluss von Hintergrundmerkmalen, Nebentätigkeit während des Studiums sowie der Studienmotivation auf den Studienabschluss ohne Stichprobenausfälle analysiert werden kann.

Dieser Ansatz wurde erstmalig am Fachbereich Wirtschaftswissenschaft der FU Berlin realisiert. Der Aufsatz beschreibt die Durchführung dieses Konzepts sowie Analyseergebnisse für den Studienverlauf und Studienabbrüche. Im Ergebnis erhalten wir, dass sich ein Studienabbruch schon in der Studieneingangsphase anhand der erworbenen Leistungspunkte und der Selbsteinschätzung der Studierenden sehr genau vorhersagen lässt. Hingegen liefern die Schulnote und die sozialen Hintergrundmerkmale keine zusätzliche Information für einen Studienabbruch.

Stichworte: Prospektives Erhebungsdesign, Längsschnitt-Analyse

Zusammenfassung: Datengetriebene Transformationen im Kontext kleinräumiger Schätzmethoden

Modelle für kleinräumige Schätzmethoden hängen typischerweise von der Erfüllung von Modellannahmen ab. So benötigte eine häufig benutzte Version des empirisch besten Prädiktors normalverteilte Fehlerterme im zugrundeliegenden gemischten linearen Modell, obwohl diese Eigenschaft in Anwendungen mit realen Daten nur selten gegeben ist. Das vorliegende Manuskript empfiehlt die Verwendung von skalierten datengetriebenen Transformationen im Gegensatz zu herkömmlichen ad-hoc gewählten Transformation, um mit möglichen Verletzungen der Modellannahmen umzugehen. Es werden verschiedene Transformationstypen und die Schätzung der korrespondierenden Transformationsparameter, im Kontext des gemischten linearen Modells, untersucht und umfassend analysiert. Des Weiteren werden die Transformationen in der kleinräumigen Schätzung linearer sowie nichtlinearer Indikatoren demonstriert.

Die Verwendung skalierter Transformationen ist entscheidend. Durch diese kann das gemischte lineare Modell mit Standardsoftware geschätzt werden, was wiederum die Arbeit des Datenanalysten wesentlich vereinfacht. Die Schätzung des mittleren quadratischen Fehlers wird adaptiert, um die Schätzung der Transformationsparameter zu berücksichtigen. Dabei werden sowohl parametrische als auch semi-parametrische (wild) Bootstrapverfahren untersucht. Die vorgeschlagenen Methoden werden anhand der Schätzung von Armuts- und Ungleichheitsindikatoren für die Gemeinden des mexikanischen Staates Guerrero demonstriert. Dabei werden reale Umfrage- und Zensusdaten genutzt. Sowohl Simulationsstudien als auch die Fallstudie in Guerrero zeigen, dass sorgfältig gewählte datengetriebene Transformationen kleinräumige Schätzungen verbessern können.

Keywords: Random Effects, Bootstrap, adaptive Transformationen, Maximum Likelihood, Poverty Mapping, Small Area Estimation

Zusammenfassung: Ein Framework zur Produktion von kleinräumigen Schätzern basierend auf Area-Level Modellen in R

Das R Paket **emdi** ermöglicht die Schätzung von regional disaggregierten Indikatoren durch die Anwendung von Small Area Schätzmethoden und stellt Werkzeuge zur Modellkonstruktion und -diagnose sowie zur Präsentation und für den Export der Ergebnisse zur Verfügung. Die Version 1.1.7 des Pakets beinhaltet Small Area Schätzmethoden für Individualdaten. Diese Methoden benötigen Mikrodaten, welche häufig Datenschutzbeschränkungen unterliegen und daher teilweise nur schwer oder nicht verfügbar sind. Im Vergleich hierzu sind sogenannte „Area-Level“ Methoden weniger anspruchsvoll im Hinblick auf (a) die Verfügbarkeit von Daten, da die Modelle auf aggregierten Daten basieren und (b) die verfügbare Rechenleistung. Des Weiteren ermöglichen „ Area-Level“ Methoden es, die Eigenschaften des Erhebungsdesigns auszunutzen. Aus diesem Grund wurden das „Area-Level“ Modell (Fay und Herriot, 1979) und vielfältige Erweiterungen in Version 2.0.1 des Paketes **emdi** ergänzt. Unter anderem beinhalten diese Erweiterungen (a) transformierte „Area-Level“ Modelle mit Rücktransformation, (b) räumliche und robuste Modelle, (c) adjustierte Varianzschätzer und (d) „Area-Level“ Modelle, welche den Messfehler berücksichtigen. Die zugehörigen Schätzer für die mittleren quadratischen Fehler sind ebenfalls enthalten und ermöglichen eine Bewertung der Unsicherheit der Prädiktion. Um das Paket so anwenderfreundlich wie möglich zu gestalten, enthält es zudem nützliche Werkzeuge zur schrittweisen Variablenselektion und Modelldiagnostik. Funktionen zum Benchmarking, zur Erstellung qualitativ hochwertiger Karten und zum Ergebnisexport sind ebenfalls enthalten. Somit ermöglicht **emdi** den vollständigen Analyse-Prozess - von der Modellkonstruktion über Diagnostik bis zur Präsentation und zum Export. Die Funktionalitäten des Pakets werden anhand von synthetischen Daten zu österreichischen Gemeinden illustriert.

Keywords: Fay-Herriot, offizielle Statistik, Survey Statistik, Small Area Estimation

Zusammenfassung: ammlogit: Schätzung und Prädiktion mithilfe eines gemischten multinomialen Logit Modells unter Verwendung aggregierter Daten

Das R-Paket **ammlogit** beinhaltet die Implementierung einer gemischten multinomialen logistischen Regression für aggregierte Daten (Molina, Saei u. a., 2007). In diesem Manuskript wird die ursprüngliche Methode erweitert, um Prädiktionen für out-of-sample Regionen zu ermöglichen. Außerdem ermöglicht die Adaption eine Nutzung von unterschiedliche Informationen über die Kovariaten in der Stichprobe und der Population. Über die Modellschätzung hinaus bietet **ammlogit** Unterstützung in der Modellbewertung und stellt R-typische Methoden für die geschätzten Modelle zur Verfügung. Die Prädiktion erfolgt automatisch sowohl für die Anteilswerte als auch für die absoluten Anzahlen. Es werden sowohl Schätzer für die Regionen als auch Schätzer für die vorliegenden Subgruppen bestimmt. Ein parametrischer Bootstrap ist implementiert, um die mittlere quadratische Abweichung jeder dieser Prädiktionen zu schätzen. Die Ergebnisvisualisierung der Punkt- und Abweichungsschätzer wird ebenfalls durch **ammlogit** ermöglicht. Sofern passende Shapedaten vorliegen, können mit **ammlogit** qualitativ hochwertige Choroplethenkarten erstellt werden.

Keywords: R, gemischte multinomiale logistische Regression, Small Area Estimation, Poverty Mapping

Zusammenfassung: Asymptotische Verteilung von Regressionsquantilen in einem gemischten Modell

Lineare Quantilmodelle erlauben eine robuste Analyse der bedingten Verteilung einer Zielvariable. Die Erweiterung um einen Fehlerterm, welcher gruppierte Zufallseffekte erfasst, erweitert den möglichen Einsatzbereich um Anwendungen mit einer komplexeren Abhängigkeitsstruktur, wie sie in vielen Studien auftritt. In diesem Manuskript schlagen wir ein verbessertes theoretisches Verständnis von linearen gemischten Quantilmodellen vor, indem wir das asymptotische Verhalten des zugehörigen Maximum-Likelihood-Schätzers zeigen. Wir beweisen die Konsistenz des Schätzers und zeigen seine asymptotische Normalverteilung. Zusätzlich leiten wir einen Plug-In Varianzschätzer her und demonstrieren sein Verhalten in endlichen Stichproben in einer Simulationsstudie.

Keywords: Quantilregression, gemischtes lineares Modell, Maximum Likelihood, asymmetrische Laplace Verteilung

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.

Göttingen, February, 2021

Sören Pannier