# Causality, Prediction, and Replicability in Applied Statistics: Advanced Models and Practices

Dissertation zur Erlangung des Doktorgrades an der
Wirtschaftswissenschaftlichen Fakultät der
Georg-August-Universität Göttingen

vorgelegt von
**Peter Pütz**
aus Minden

Göttingen, 2019

# Prüfungskommission

Erstgutachter: Prof. Dr. Thomas Kneib
Zweitgutachter: Prof. Dr. Sebastian Vollmer
Drittprüfer: Prof. Dr. Bernhard Brümmer

# „Wer eine Jogginghose trägt, ...“

„... hat die Kontrolle über sein Leben verloren“, so sagte einst der kürzlich verstorbene Modeschöpfer Karl Lagerfeld. Das mag sein. Ohne Jogginghose hätte ich diese Doktorarbeit jedoch nicht schreiben können. Ebenso wenig hätte ich es ohne eine Vielzahl an Menschen geschafft, denen ich im Folgenden danken möchte.

Als erstes bedanke ich mich bei meinen Eltern für die stetige Unterstützung.

Ich bedanke mich bei Thomas Kneib für die Einstellung an der Uni Göttingen als statistischer Berater für Studis und die spätere „Versetzung“ in den SFB 990. Ich danke Dir für unverwüstlich gute Laune, Ermutigung, Ehrlichkeit, Rat aller Art, Verständnis, gute Witze und die vielen Freiheiten und Möglichkeiten, die ich bekommen habe. Darunter fallen insbesondere die Forschungsaufenthalte in Indien, Indonesien und Portugal. Insgesamt habe ich selten einen Menschen kennengelernt, von dem man fachlich und menschlich so viel lernen kann. Danke, Thomas.

Ich danke Sebastian Vollmer für die Zweitbetreuung meiner Arbeit und die Möglichkeit, an dem spannenden Forschungsprojekt in Bihar teilnehmen zu können. Die Erfahrungen aus dieser völlig verrückten Zeit dort möchte ich nicht missen. Ein großes Dankeschön geht an Bernhard Brümmer für die Drittbetreuung meiner Arbeit, um die ich recht kurzfristig gebeten habe. Vielen Dank auch für die nette Zusammenarbeit im SFB 990.

Ich bedanke mich ganz herzlich bei der Statoek-Arbeitsgruppe (fühlt Euch alle angesprochen!) für eine überragende Zeit mit Spiel, Spaß und Wissen. Unter die Arme greifen anstatt die Ellenbogen ausfahren, das beschreibt die fachliche Zusammenarbeit ganz gut, denke ich. Außerdem hätte ich nicht gedacht, dass ein Haufen promovierender und promovierter Statistik-Menschen so witzig und menschlich sein kann. Ein besonderer Dank geht an Simone Maxand und Hauke Thaden für den guten Start in Göttingen sowie Maike Hohberg und Stephan Bruns für die inspirierende und spaßige Zusammenarbeit bei zwei Papieren.

Ich danke all den coolen Menschen vom SFB 990, von denen ich sehr viel über Dinge lernen konnte, die nicht mit Statistik zusammenhängen, aber umso wichtiger sind.

Zusätzlich bedanke ich mich herzlich bei Dörte Dede, Ivonne Hein, Ivonne Hofmann und Barbara Wick für ausgesprochene Freundlichkeit und die Hilfe bei zahlreichen Problemen.

Ein großer Dank für das Lesen der Doktorarbeit und die vielen guten Verbesserungsvorschläge geht an Stephan Bruns, Franziska Dorn, Maike Hohberg, Marion Krämer, Bruno Santos, Benjamin Säfken, Paul Schneider, Alexander Sohn, Elmar Spiegel und Mareike Stücken.

Herzlichen Dank Dir, Marion, auch dafür, dass ich Indien Dein Datenknecht sein durfte. Es war eine herausragende Zeit dort und ich würde immer wieder Dein Angestellter sein wollen.

Abschließend bleibt zu bemerken, dass Aufwand und Ertrag bzw. Güte dieser Doktorarbeit in keinem angemessenen Verhältnis stehen. Die Zeit, die ich für das Schreiben gebraucht habe, hätte ich anders sicher sinnvoller und effizienter nutzen können. Genauso sicher ist für mich jedoch, dass diese Zeit kaum hätte schöner sein können. Dafür danke ich auch allen Menschen, die ich auf dieser Seite nicht namentlich erwähnt habe, die sich aber zurecht angesprochen fühlen.

# Abstract

Statistical tools to analyze research data are widely applied in many scientific disciplines and the need for adequate statistical models and sound statistical analyses is apparent. This thesis addresses limitations in statistical models commonly used to identify causal effects and for prediction purposes. Moreover, difficulties in the replicability of statistical results are revealed and remedies are suggested.

With regard to causality, the incorporation of penalized splines into fixed effects panel data models is proposed. Fixed effects panel data models are often used in order to establish causal effects since they control for unobserved time-invariant heterogeneity of the study entities. The inclusion of penalized splines relieves the researcher from determining functional shapes of the covariate effects. Instead, the functional forms are allowed to be flexible and are estimated based on the data at hand such that a data-driven degree of nonlinearity is identified. Simultaneous confidence bands are presented as a computationally fast and reliable uncertainty measure for the estimated functions. Furthermore, this thesis studies causal effects not only on the expectation but on all aspects of the distribution of the dependent variable. In particular, generalized additive models for location, scale and shape are introduced to (quasi-)experimental methods. A step-by-step guide demonstrates how the proposed methodology may be applied and provides insights which may go unnoticed in common regression frameworks.

In the domain of prediction, a small area prediction problem is considered. It is shown how to obtain reliable up-to-date welfare estimates when an outdated census without information on income and a more recent survey with information on income are available. Instead of using survey variables to explain income in the survey, the proposed approach uses variables constructed from the census. The underlying assumptions are less restrictive than those in commonly applied methods in this field that are tailored to situations with simultaneous census and survey collection.

As an overarching topic relating to all statistical analyses, the replicability of statistical results is considered from two viewpoints. On the one hand, the prevalence of reporting errors in statistical results is investigated. On the other hand, studies are replicated if possible by using the same data and software code as in the reference study. It is shown that replicability is frequently made impossible by reporting errors as well as by missing data and software code. At the same time, simple solutions to enhance replicability in future research are presented. Open data and software code policies together with a vivid replication culture seem to be most promising.

# Zusammenfassung

Statistische Methoden zur Analyse von Forschungsdaten werden in vielen wissenschaftlichen Disziplinen eingesetzt. Der Bedarf an adäquaten statistischen Modellen und fundierten statistischen Analysen ist offensichtlich. Diese Dissertation adressiert Einschränkungen in statistischen Modellen, die üblicherweise zur Ermittlung kausaler Effekte und zu Vorhersagezwecken verwendet werden. Darüber hinaus werden Probleme hinsichtlich der Replizierbarkeit statistischer Ergebnisse aufgedeckt und Lösungen vorgeschlagen.

Im Hinblick auf Kausalität wird die Integration von pönalisierten Splines in Paneldatenmodelle mit fixen Effekten vorgeschlagen. Diese Modelle werden häufig zur Ermittlung kausaler Effekte verwendet, da sie für nicht beobachtete zeitinvariante Heterogenität der Beobachtungseinheiten kontrollieren. Die Einbeziehung von pönalisierten Splines befreit die Forscherin von der Aufgabe, die funktionalen Formen der Effekte der Kovariaten selbst festzulegen. Stattdessen dürfen die Funktionsformen flexibel sein und werden anhand der vorliegenden Daten geschätzt, sodass ein datengetriebenes Maß an Nichtlinearität bestimmt wird. Als eine rechenunaufwendige und zuverlässige Methode zur Unsicherheitsmessung für die geschätzten Funktionen werden simultane Konfidenzbänder vorgestellt. Darüber hinaus untersucht diese Arbeit kausale Effekte nicht nur auf den Erwartungswert, sondern auf alle Aspekte der Verteilung der abhängigen Variablen. Insbesondere werden generalisierte additive Modelle für Lokation, Skala und Form mit (quasi-)experimentelle Methoden verbunden. Eine Schritt-für-Schritt-Anleitung zeigt, wie die vorgeschlagene Methodik angewendet werden kann und Einblicke liefert, die in herkömmlichen Regressionsmodellen unbemerkt bleiben könnten.

Im Bereich der Prädiktion wird ein Problem der Vorhersage kleinräumiger Daten betrachtet. Es wird gezeigt, wie verlässliche und aktuelle Wohlfahrtsschätzungen erhalten werden können, wenn ein veralteter Zensus ohne Informationen über das Einkommen und neuere Surveydaten mit Informationen über das Einkommen verfügbar sind. Anstelle der Nutzung von Variablen aus dem Survey zur Vorhersage von Einkommen verwendet der vorgeschlagene Ansatz aus dem Zensus konstruierte Variablen. Die dafür notwendigen Annahmen sind weniger einschränkend als die in gewöhnlich verwendeten Verfahren, die auf Situationen mit gleichzeitiger Erhebung von Zensus und Survey zugeschnitten sind.

Als übergreifendes Thema aller statistischen Analysen wird die Replizierbarkeit statistischer Ergebnisse aus zwei Blickwinkeln betrachtet. Zum einen wird die Häufigkeit von Berichtsfehlern in statistischen Ergebnissen untersucht. Auf der anderen Seite wird versucht, Studien unter Verwendung der gleichen Daten und des gleichen Softwarecodes zu replizieren. Es wird gezeigt, dass die Replizierbarkeit häufig durch Berichtsfehler sowie durch fehlende Daten und Softwarecode unmöglich gemacht wird. Gleichzeitg werden einfache Lösungen zur Verbesserung der Replizierbarkeit in zukünftiger Forschung präsentiert. Vorschriften zur Offenlegung von Daten und Softwarecode zusammen mit einer regen Replikationskultur scheinen die vielversprechendsten zu sein.

# Contents

# 1 Introduction

Many scientific disciplines such as biology, economics, epidemiology, medicine, physics, psychology, and sociology rely on statistical analyses of research data. Handbooks on applied statistics are numerous and the need to properly apply statistical methods to the rising amount of available data is apparent. In top economic journals, for instance, the share of papers using some sort of statistical data analysis has increased from less than 50% to more than 70% from 1963 to 2011 (Hamermesh, 2013). Due to ongoing advances in computing power and boosting storage capacities, the availability of data is likely to rise further and thus future research can be expected to have a strong quantitative component (e.g., Bello-Orgaz et al., 2016; Erevelles et al., 2016; Stephens et al., 2015; Einav and Levin, 2014).

Statistical analyses often imply the application of models. In this thesis, the focus is on regression models that are among the most frequently used statistical modeling tools in the above-mentioned disciplines. Regression models describe the relationship of a dependent variable and one or more explanatory variables. A basic regression model is the linear additive model

$$y = \beta_0 + \sum_{k=1}^{K} \beta_k x_k + u, \tag{1.1}$$

where $y$ is the dependent variable (response), $x_1, \ldots, x_K$ are explanatory variables (covariates) with associated regression coefficients $\beta_1, \ldots, \beta_K$, while $\beta_0$ is an intercept and $u$ is an error term.

Regression models usually involve restrictive assumptions such as independent and identically distributed errors or linearity in the parameters as implied by Equation (1.1). The conditional expectation of the dependent variable is in the vast majority of cases the sole subject of analysis, although other aspects of the distribution of the dependent variable may also be of interest. Thus, regression models may answer specific questions approximately under specific conditions but require scrutiny and cautious interpretation whenever applied to a particular dataset.

Regression models can be used for estimating causal relationships, for instance in program evaluation, as well as for prediction purposes, for example when forecasting time series. Sections 1.1 and 1.2 elaborate on these two goals of regression analysis from a general viewpoint. Irrespective of the goal of a regression analysis, its results are supposed to be replicable. Since parts of this thesis deal with the replicability of statistical results, the underlying concept of replicability is presented in Section 1.3.

This thesis includes four scientific articles that are summarized in Section 1.4. It is described how these articles classify thematically into the domains presented in Sections 1.1, 1.2, and 1.3 and how they address related challenges and limitations. Additionally, the authors' contributions to the respective articles are listed. The four articles are printed in full length in Chapters 2, 3, 4, and 5, respectively. Chapter 6 concludes and draws avenues for future research. While this thesis focuses on economic applications, the arguments carry over directly to the other disciplines mentioned above.

## 1.1 Causality

A causal effect of a selected explanatory on a dependent variable is an effect that can only be attributed to a change in this selected explanatory variable. To put it differently, causality means that a specific action induces a specific measurable consequence (Stock and Watson, 2011, p. 6). One conceptually

simple approach in order to establish a causal effect in a study is to conduct a randomized controlled trial. The study participants are randomly assigned to one or more treatment groups or to a control group with the treatment groups receiving a treatment or intervention and the control group not. If the sample size is large, the random assignment most likely ensures that the only systematic difference between the treatment and control groups is the treatment. In the simplest case with a dummy treatment variable and no further covariates, Equation (1.1) simplifies to

$$y = \beta_0 + \beta_1 x + u, \tag{1.2}$$

with $x = 0$ for the control group and $x = 1$ for the treatment group. For a given dataset, an estimate for the causal effect on the expected value of the response variable may then be obtained by the ordinary least squares estimate for $\beta_1$, which is equivalent to the mean difference between treatment group and control group with respect to the dependent variable. In principal, effects on other quantities of the groups' response distributions, for example the variance, could be estimated as well.

Randomized controlled trials are widely applied in many research fields. One prominent example is medicine with its numerous clinical trials. Also in economics, randomized controlled trials have been identified as a valid method to evaluate programs (e.g., Angrist and Pischke, 2010). In Chapter 3, the famous poverty alleviation program Progresa conducted in Mexico is described. It was implemented as a randomized controlled trial on a large scale and mirrors the growing importance of rigorous program evaluation in development economics as shown by Cameron et al. (2016).

The execution of randomized controlled trials is often prevented by the impracticability of the randomization due to ethical or political reasons, high costs or time constraints (Bärnighausen et al., 2017). If one nonetheless needs to identify a causal effect, quasi-experimental studies can be similarly fruitful. Their central idea is to exploit specific study designs and data structures such that the assignment of the study participants to treatment and control groups can be considered "as if" it was random (Stock and Watson, 2011, ch. 13). Quasi-experimental studies are regularly applied in many research fields, for instance in economics. According to Angrist and Pischke (2010), better research designs including the sound usage of quasi-experimental designs have led to a "credibility revolution" in economics. Empirical economists nowadays have to defend the research design chosen to identify causal effects in their publications. Both outright randomization and "as if" randomization of the treatment may be convincing arguments for this purposed. Quasi-experimental studies are also popular in other disciplines where outright randomization is often not feasible, for example in epidemiology (e.g., Bärnighausen et al., 2017, and the whole corresponding issue in the *Journal of Clinical Epidemiology* on this topic).

Regression models applied in order to establish causal effects are often not as trivial as in Equations (1.1) and (1.2). They may involve many potentially interacting covariates, dependence between observations and nonlinear treatment effects if the treatment is measured on a continuous scale. Angrist and Pischke (2008) nonetheless call their book about (quasi-)experimental methods in economics "Mostly Harmless Econometrics" and argue that advanced statistical techniques are typically unnecessary when the focus lies on causality. This statement will be put into question in this thesis.

## 1.2 Prediction

In many studies, it is not intended to establish causal or direct effects of specific variables in a regression model. Regarding model (1.2), causality of the effect of $x$ on $y$ requires the conditional expectation of the error term given the explanatory variable to be zero, that is, $E(u|x) = 0$. In a model focusing on

prediction instead of causality, $\beta_1$ needs not to capture the causal effect of $x$ but may also incorporate the effects of unobserved variables subsumed in $u$ which are correlated with $x$. The same holds for models with several covariates as in Equation (1.1) and more sophisticated models.

While the covariate effects may also be of considerable interest in a prediction model, the main goal is to predict the dependent variable as accurately as possible for new observations with known covariate values. Similarly, an aggregate measure of the dependent variable may be the quantity of interest to predict. Prediction accuracy in this sense is usually the size of the expected prediction error, that is, the expected (squared) deviation between the values predicted by the model and the unobserved true values (e.g., Fahrmeir et al., 2013, pp. 138-139). The applied statistical methodology may comprise variable and model selection tools that aim to find a model with high explanatory power but preferably few parameters to estimate. This bias-variance trade-off associated with finding a small prediction error is often referred to in the statistics literature (e.g., Fahrmeir et al., 2013, pp. 138-148).

In general, models commonly applied for a causal analysis are arguably less flexible and easier to understand than most advanced prediction models. This is especially obvious when it comes to machine learning, a field mainly dedicated to prediction tasks. Approaches applied in machine learning such as artificial neural networks are often too complex to be written down as simple regression equations. Their estimation requires cautious specification or optimization of potentially several (hyper-)parameters and the interpretation of the effect of a single variable is awkward. Applications for machine learning techniques are manifold and range from weather forecasting to image recognition. An extensive introduction to machine learning methods including many potential applications is given, for example, by Hastie et al. (2009).

## 1.3 Replicability

Irrespective of the aim of a statistical analysis, may it be the establishment of a causal link or a good prediction, statistical results are expected to be replicable. The replicability and thus reliability of published empirical findings is at the top of the agenda for reputable research associations (e.g., National Academies of Sciences et al., 2016; Open Science Collaboration, 2015). That said, the term "replicability" is not uniquely defined across the sciences and not even within a single research field. Sometimes it is used interchangeably with "reproducibility", a term not used in the remainder of this thesis. Clemens (2017) and Goodman et al. (2016) give comprehensive discussions on the definitions of both terms. If replicability refers to whether statistical results from one study can be replicated in a second study, it is particularly controversial what "replicable" means (e.g., Open Science Collaboration, 2015; Simonsohn, 2015; Valentine et al., 2011; Cumming and Maillardet, 2006). Is it sufficient that the direction of the coefficients of interest are the same in both studies? Do the effect sizes from the second study have to be within the confidence intervals of the effect sizes of the first study? Do the respective $p$-values in both studies have to pass the same threshold?

In any case, each statistical analysis is subject to many inherently nonobjective choices by the involved researchers (e.g., Berger and Berry, 1988). These "researcher degrees of freedom" (Simmons et al., 2011) concern the data collection, sample size, study design, statistical model and reported results. Flexibility in the data analysis can go hand in hand with "questionable research practices" (John et al., 2012) such as cherry-picking findings (see also Casey et al., 2012). Such practices applied to obtain desired results, usually corresponding to statistically significant results rejecting the null hypothesis, do not only lead to a biased body of literature and erroneous study conclusions but also threat the replicability of the results in

another study (e.g., Baker, 2016). Amrhein et al. (2018) even state that "unreplicable research is mainly a crisis of overconfidence in statistical results". In their opinion, the perceived failure to replicate studies "stems from failure to recognize that statistical tests not only test hypotheses but countless assumptions and the entire environment in which research takes place. Honestly reported results must vary from replication to replication because of varying assumption violations and random variation."

In this thesis the term replicability is used in a narrower sense as it refers to replicating a study using the original data. Two types of narrow replicability are applied. The first type of replicability is considered in Chapter 5: It is checked if provided data and software code allow to obtain the same results as in the original study allowing for small, irrelevant deviations. A similar definition for replicability is used by Chang and Li (2018) and called "verification" in Clemens (2017). A potential reason for a failed replication in this sense is an erroneously reported statistical value in the published manuscript. Such a reporting error is obviously present if, for instance, the ratio of a reported coefficient estimate $\hat{\beta}_k$ for $\beta_k$ in Equation (1.1) and its associated standard error estimate $\hat{\sigma}_{\hat{\beta}_k}$ is not equal to the reported $t$-statistic, that is, $\frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}} \neq t$. Reporting errors may occur during the review or typesetting process when results or software code are not updated or due to wrongly transcribed results. Manipulating numbers to obtain desired results are another potential source. The second type of replicability referred to in Chapter 3 of this thesis is a bit broader: A study is considered as replicable if qualitatively similar results are obtained by using a sensible different statistical method applied to the same data. Clemens (2017) calls the procedure to apply altered code and model specifications to the same data a "robustness test" and more specifically, a "reanalysis test". It is obvious that statistical results may differ, even substantially, depending on the methods used. If the applied methods are sound, either the methods ask different questions or the robustness of the findings can be considered low. Questionable research practices such as reporting only the models that yield desired results, even if their assumptions are invalid, may also lead to a discrepancy.

## 1.4 Summaries of the articles

### 1.4.1 A penalized spline estimator for fixed effects panel data models

**Pütz, P. and Kneib, T.** (2018). A penalized spline estimator for fixed effects panel data models. Published in *AStA Advances in Statistical Analysis*, Vol. 102(2), 145-166.
DOI: 10.1007/s10182-017-0296-1.
The article is printed in full length in Chapter 2.

#### Content

Nonparametric and semiparametric regression methods are extremely popular in statistics when studying the impact of one or more continuous covariates on a response variable. Their main advantage is that they do not impose strong prior assumptions on the functional shape of the covariate effects but rather let the data speak for themselves such that a data-driven degree of nonlinearity is identified. Penalized splines (Eilers and Marx, 1996) offer one possibility to estimate nonlinear effects of continuous covariates. Their use is well established for regressions with cross-sectional data as well as for panel data regressions with random effects. However, when utilizing a random effects specification for panel data, one assumes that the random effects and the regression covariates are independent. Fixed effects specifications loosen this crucial assumption. In this paper, we consider fixed effects instead of random effects panel data models

and develop a first-difference approach for the inclusion of penalized splines in this case. We take the resulting dependence structure into account and adapt the construction of uncertainty measures accordingly. The latter are based on simultaneous confidence bands that provide a simultaneous uncertainty assessment for the whole estimated functions. To construct the confidence bands, we build on the ideas of Wiesenfarth et al. (2012) and exploit the mixed model representation of penalized splines. By doing so, computationally demanding resampling techniques are avoided and a fast way of inference is provided. In addition, the penalized spline estimates as well as the confidence bands are also developed for the derivatives of the estimated effects which are of considerable interest in many application areas. As an empirical illustration, we analyze the dynamics of life satisfaction over the life span based on data from the German Socio-Economic Panel (SOEP). An open-source software implementation of our methods is available in the `R` package `pamfe`.

### Classification

As fixed effects panel data models control for time-invariant unobserved heterogeneity of the study entities and allow for correlations between the fixed effects and further covariates, they are often used to identify causal effects. The main contribution of the paper is in the domain of causality (Section 1.1) as these models are combined with penalized splines. By allowing flexibility in the specification of covariate effects, may it be for variables of interest or control variables, this extension is less restrictive than the commonly applied parametric fixed effects panel data models.

### Contributions of the authors

I designed the overall structure of the paper, wrote the draft for the whole paper and was responsible for alterations made. I designed and realized the simulation study and implemented the proposed methods in the `R` package `pamfe`. I prepared the SOEP data and conducted all data analyses in `R` for the application of the methods. Thomas Kneib developed the research question. He supported the development of the paper by proofreading and providing statistical insights. Additionally, he improved the paper by reformulating the introduction.

## 1.4.2 Treatment effects beyond the mean using GAMLSS

**Hohberg, M., Pütz, P. and Kneib, T.** (2019). Treatment effects beyond the mean using GAMLSS. Submitted to *The American Statistician*, under review since December 28, 2018. Available on: `https://arxiv.org/pdf/1806.09386.pdf`.
The article is printed in full length in Chapter 3.

### Content

Program evaluation in economics typically identifies the effect of a policy or an intervention as the average difference between treatment and control group with respect to the response variable. Concentrating on mean differences between treatment group and control group is likely to miss important information about changes along the whole distribution of an outcome, for example in terms of an unintended increase in inequality, or when targeting ex ante vulnerability to a specific risk. These concepts rely on additional measures such as the variance and skewness of the response. For a systematic and coherent

analysis of treatment effects on all parameters of the response distribution, we introduce generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005, GAMLSS), to the program evaluation literature. We provide practical guidance on the usage of GAMLSS by reanalyzing data from the Mexican Progresa program. The results are very similar to the original study by Angelucci and De Giorgi (2009) when looking at mean differences in consumption between treatment and control group. Using GAMLSS furthermore allows us to investigate effects beyond the mean not considered in the original study. Contrary to expectations, we find no significant effect of a cash transfer on the conditional inequality level between treatment and control group. This practical example considers only the case of a randomized controlled trial. Popular quasi-experimental methods in the context of establishing causal effects include regression discontinuity designs, differences-in-differences, panel data techniques, and instrumental variable regression. In this paper, we develop frameworks for combining each of these methods with GAMLSS.

### Classification

The paper considers causal effects as introduced in Section 1.1 but extends the common focus on the conditional expectation of the dependent variable. An advanced statistical method, namely GAMLSS, is combined with common approaches to identify causal effects. Additionally, the paper adds to the replicability of science (Section 1.3) as a popular study is replicated by using this advanced method.

### Contributions of the authors

I wrote the step-by-step guide on how to apply GAMLSS and conducted the corresponding data analyses in `R` for the application of GAMLSS to the Progresa data. Excluding the paragraph on instrumental variables and GAMLSS, I was responsible for the whole appendix. This included the combination of GAMLSS with program evaluation methods and the elaborations on bootstrap inference. I worked in constant exchange with my co-authors on the structure and content of the whole paper. Maike Hohberg had the research idea and designed the overall concept of the paper. She wrote the main part of the main body of the text and the paragraph on the combination of GAMLSS and instrumental variable regression in the appendix. Thomas Kneib supported the development of the paper by proofreading and statistical discussions.

## 1.4.3 Small area estimation of poverty under structural change

**Lange, S., Pape, U. J. and Pütz, P.** (2018). Small area estimation of poverty under structural change. Policy research working paper 8472, the World Bank. Available on:
`http://documents.worldbank.org/curated/en/612621528823563131/pdf/WPS8472.pdf`.
The article is printed in full length in Chapter 4.

### Content

Small area poverty and wealth maps allow the design of policies dependent on spatial differences in welfare, for instance the allocation of financial aid. While such a map is useful for policy makers and researchers when small geographic units (e.g., villages) are discernable, welfare estimates based on household surveys are typically not representative at such levels of disaggregation. On the other hand, most censuses do

not contain information on income or consumption expenditures required to calculate (financial) welfare. The most frequently applied welfare mapping approaches rely on combining contemporaneously collected survey and census information. Typically, a regression model to explain income is estimated based on survey data in the first step. In the second step, the resulting estimates are used to predict income for all census households. This methodology requires commonality assumptions on the explanatory variables in both data sources which hardly hold if the datasets are not collected simultaneously. While the monitoring of wealth over time and the generation of up-to-date wealth maps are of eminent interest to economists and policy makers, little attention has been paid to common situations in which considerable time has passed between census and survey collection. In this contribution, we present a new approach which allows the generation of up-to-date poverty maps when an outdated census and a more recent survey are available. Instead of using survey variables to explain income in the first place, the new approach uses outdated census information to explain income values in the survey. The proposed technique has lower data requirements and weaker assumptions than common small area poverty estimators. Applications to simulated data and to poverty estimation in Brazil show an overall good performance. Thus, our approach is a promising tool to generate reliable and up-to-date welfare estimates in many situations with an outdated census and a more recent survey. Furthermore, the method is applicable to a wide range of outcome measures and research questions beyond welfare mapping.

### Classification

The estimated regression models considered in this paper solely focus on prediction as described in Section 1.2. In particular, the first step of the procedure requires to find a model which predicts household income values as precisely as possible. The quantity of interest is an aggregate measure of income values for all households in a specific area. For most or all of these households, the income is not observed but has to be predicted based on the parameter estimates from the first step. The ciritical point is the unavailability of current covariate values for these households. To this end, the contribution of the paper does not lie in finding a sophisticated prediction model but in finding a prediction model that allows to make accurate predictions for all households for which only outdated census information is available.

### Contributions of the authors

I designed the overall structure of the paper and wrote the draft for the whole paper. I derived the properties of the proposed estimator and designed the simulation study and implemented it in `Stata`. Additionally, I conducted all data analyses in `Stata` for the application of the method to the data from Minas Gerais. Simon Lange and Utz Johann Pape developed the research question, supported the development of the paper by proofreading and discussing the overall idea, structure and content of the paper. Additionally, they improved the readability of the paper substantially by reformulating specific parts.

## 1.4.4 The (non-)significance of reporting errors in economics: Evidence from three top journals

**Pütz, P. and Bruns, S. B.** (2019). The (non-)significance of reporting errors in economics: Evidence from three top journals. Submitted to *The Review of Economics and Statistics*, under review since January 31, 2019.
The article is printed in full length in Chapter 5.

## Content

Statistical information such as coefficients, standard errors, test statistics, and $p$-values constitute the core output in empirical economics and quantitative research in general. These statistical values are essential for the cumulative research process and frequently used in evidence-based decision making. Therefore, it is of eminent importance that they are reported correctly. Erroneous statistical information undermines the reliability of published findings and questions the quality of the peer-review process in academia. In this paper, we investigate the prevalence and drivers of reporting errors in three top economic journals (American Economic Review, Quarterly Journal of Economics and Journal of Political Economy) between 2005 and 2011. Reporting errors are defined as inconsistencies between reported statistical values such as coefficients and standard errors on the one hand and significance levels labeled by eye-catchers (usually asterisks) on the other hand. Our dataset comprises 370 articles with 30,993 tests of central hypotheses mostly from regression tables. We find that 34% of all articles contain at least one reporting error and 19% contain at least one strong reporting error that renders a statistically significant finding non-significant or *vice versa*. The rate of errors at the test level is very small: Only 1.3% of all hypotheses tests are afflicted by a reporting error and 0.5% by a strong reporting error. A survey sent to all authors in our dataset whose articles included at least one error and systematic replications shed light on potential sources. Errors seem to occur frequently in the eye-catchers and by manually transferring findings from statistical software to word-processing programs. Moreover, regression analyses suggest that error rates differ between journals which may be related to differences in the transparency guidelines and the quality of type setting. Our findings imply easy remedies to reduce the rate of reporting errors in future research, such as applying automated algorithms to check the consistency of statistical information before publication or more generally to ban eye-catchers and the corresponding dichotomization into statistically significant or non-significant findings. Open data and software code policies in line with a vivid replication culture seem to be equally promising remedies.

## Classification

This paper focuses on very narrow replicability or "verification" (Clemens, 2017) as described in Section 1.3. Two different approaches used in the paper can be attributed to this topic. First, reporting errors in published statistical results are identified. A reporting error in a statistical result immediately implies that the afflicted result cannot be replicated. Second, 64 articles are tried to be replicated by using the original data and code if available.

## Contributions of the authors

I wrote the draft for the whole paper. I realized all data coding and management tasks and conducted all replications. Additionally, I was responsible for all empirical analyses and most of their implementation in `R`. Stephan B. Bruns and I jointly developed the research idea. Stephan B. Bruns assisted with the implementation of the algorithm to detect reporting errors in `R`. He improved the readability and content of the paper substantially by proofreading and revising the structure and all paragraphs.

# 2 A penalized spline estimator for fixed effects panel data models

# A penalized spline estimator for fixed effects panel data models

Peter Pütz*, Thomas Kneib†

**Abstract**

Estimating nonlinear effects of continuous covariates by penalized splines is well established for regressions with cross-sectional data as well as for panel data regressions with random effects. Penalized splines are particularly advantageous since they enable both the estimation of unknown nonlinear covariate effects and inferential statements about these effects. The latter are based, for example, on simultaneous confidence bands that provide a simultaneous uncertainty assessment for the whole estimated functions. In this paper, we consider fixed effects panel data models instead of random effects specifications and develop a first-difference approach for the inclusion of penalized splines in this case. We take the resulting dependence structure into account and adapt the construction of simultaneous confidence bands accordingly. In addition, the penalized spline estimates as well as the confidence bands are also made available for derivatives of the estimated effects which are of considerable interest in many application areas. As an empirical illustration, we analyze the dynamics of life satisfaction over the life span based on data from the German Socio-Economic Panel (SOEP). An open source software implementation of our methods is available in the `R` package `pamfe`.

**Keywords**: First-difference estimator; Life satisfaction; Panel data;Penalized splines; Simultaneous confidence bands

*University of Göttingen, Faculty of Economic Sciences, Chair of Statistics, e-mail: ppuetz@uni-goettingen.de.
†University of Göttingen, Faculty of Economic Sciences, Chair of Statistics.
An online supplement and the `R` package `pamfe` can be found at `https://www.uni-goettingen.de/de//511092.html`.

## 2.1 Introduction

Nonparametric and semiparametric regression methods are extremely popular in statistics and econometrics when studying the impact of one or more continuous covariates on a response variable. Their main advantage is that they do not impose strong prior assumptions on the functional shape of the covariate effects but rather let the data speak for themselves such that a data-driven amount of nonlinearity is identified. In this paper, our interest lies in estimating regression models with flexible covariate effects for panel data. We therefore think of $N$ persons observed at $T$ points in time and consider an additive panel data model of the form

$$y_{it} = \gamma_i + \sum_{h=1}^{H} f_h(x_{hit}) + u_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T,$$

where $y_{it}$ is the response variable of interest, $f_1(x_{1it}), \ldots, f_h(x_{hit})$ are some unknown smooth functions representing the potentially nonlinear effects of $H$ continuous covariates, $u_{it}$ are independent and identically distributed normal error terms with constant variance and $\gamma_i$ are individual-specific, time-invariant effects either allowed (fixed effects model) or not allowed (random effects model) to be correlated with the covariates. For the specification of the covariate effects $f_1(x_{1it}), \ldots, f_h(x_{hit})$, we rely on penalized B-splines (Eilers and Marx, 1996) which approximate a potentially nonlinear effect of interest by a rich B-spline basis while adding a penalty to the penalized least squares criterion to regularize estimation. In addition to their computational attractiveness, penalized splines are also easily combined with parametric effects to obtain semiparametric partially linear models and allow for easy access to uncertainty measures.

So far, penalized splines have mostly been used for either cross-sectional data or in combination with random effects specifications for panel data. The main reason for this is the fact that the penalty considered for penalized splines fits nicely together with the "penalty" imposed by the random effects and in fact penalized splines can be considered a special type of random effects model as well, see, for example, Ruppert and Wand (2003) or Fahrmeir et al. (2013). However, when utilizing a random effects specification for panel data, one has to critically evaluate whether correlations between the random effects and the regression covariates are present. Fixed effects specifications loosen this crucial assumption and are particularly popular in econometrics. To avoid the incidental parameter problem that arises when including fixed effects, estimation is then typically based on first-order differenced or demeaned data. For nonparametric and semiparametric panel data models with fixed effects, a growing strand of literature has emerged during the last years, including Baltagi and Li (2002), Su and Ullah (2006), Henderson et al. (2008), Mammen et al. (2009), Qian and Wang (2012), Zhang et al. (2011) and Chen, Gao and Li (2013). Comprehensive literature reviews are provided by Su and Ullah (2011) and Chen, Li and Gao (2013). While having different concepts to handle the fixed effects and strictly parametric effects, all discussed methods have in common that they rely on some kind of kernel estimator for the nonparametric model components. Simultaneous confidence bands for kernel estimators have been discussed extensively for cross-sectional data, see, for instance, Eubank and Speckman (1993), Neumann and Polzehl (1998), Claeskens and Van Keilegom (2003) and Härdle et al. (2004). Furthermore, confidence bands for polynomial spline estimators have been discussed, among others, by Yang (2008) and Wang and Yang (2009), while the most recent literature on Bayesian confidence bands (or credible bands) comprises Crainiceanu et al. (2007) and Krivobokova et al. (2010). An attractive alternative to construct simultaneous confidence bands for a broad class of unbiased nonparametric regression estimators is shown in Sun and Loader (1994). In particular, they exploit the volume-of-tube formula (Weyl, 1939) to determine the tail probability of suprema of Gaussian random processes. Krivobokova et al. (2010) use the same ideas and the link

between penalized splines and mixed model to construct simultaneous confidence bands for univariate penalized spline estimators. Their confidence bands are computable with acceptable computational effort and exhibit excellent properties even in fairly small samples. The extension to complex additive models can be found in Wiesenfarth et al. (2012). Apart from the rich literature for cross-sectional data, recent work by Li et al. (2013) pioneered in the field of uncertainty assessments for the above-mentioned kernel-based fixed effects panel data estimators. Since Li et al. (2013) thereby rely on bootstrapping techniques, inferences on nonlinear model parts become challenging or at least computationally demanding in cases of large sample sizes and many nonparametric model components.

To overcome this difficulty, we consider a penalized spline specification for the nonparametric model components and apply first-order differences to the model. This basically implies a differenced basis function approximation of the nonparametric effects while relying on the same parameterization of the penalized spline as the original model. To account for the serial correlation induced by first differencing, we use a generalized least squares (GLS) criterion. Utilizing the mixed model representation of penalized splines, we develop a fast way of inference for first-difference penalized spline estimates via simultaneous confidence bands building on the ideas of Wiesenfarth et al. (2012) for cross-sectional data. This also allows us to derive simultaneous confidence bands for the derivatives of the nonparametric effects.

To illustrate the applicability of our methods, we use the information from the German Socio-Economic Panel (SOEP) database[1] on the dynamics of life satisfaction over the life span. So far, no consensus on the functional form of the relationship between age and life satisfaction has been reached. Typically, it is modeled via a strictly parametric specification, which might be too restrictive and is therefore likely to affect the results adversely. Our more flexible approach avoids this issue and also accounts for individual heterogeneity among the survey respondents by including fixed effects.

In terms of the model specification, our approach is closely related to Hajargasht (2009) who also proposed a penalized spline estimator for fixed effects panel data, based on the within-transformation, that is, demeaned data. However, our approach differs from the one by Hajargasht (2009) with respect to the following important aspects: (i) we use the mixed model representation of penalized splines not only to obtain a data-driven estimate for the smoothing parameter but also simultaneous confidence bands, (ii) we develop and investigate inferences for the potentially nonlinear effects directly and for the derivatives, (iii) we provide an open source implementation in the accompanying `R` package `pamfe` that enables practitioners to apply the proposed method, and (iv) we apply our approach to real world data in a complex semiparametric model with multiple nonparametric components.

The remainder of this paper is organized as follows: First-difference penalized spline estimation for panel data models is introduced in Section 2.2. Inference via simultaneous confidence bands is considered in Section 2.3. In Section 2.4, the performance of our approach is tested in a simulation study while the empirical investigation of the dynamics of life satisfaction is described in Section 2.5. Section 2.6 summarizes our conclusions and discusses directions for future research.

[1] Socio-Economic Panel (SOEP), data of the years 1984-2011, version 28, SOEP, 2012, doi: 10.5684/soep.v28.

## 2.2 Penalized splines for cross-sectional and panel data

### 2.2.1 Penalized splines in the cross-sectional context

We start our considerations by discussing penalized spline specifications for cross-sectional data. Consider the additive regression model

$$y_i = \beta_0 + \sum_{h=1}^{H} f_h\left(x_{hi}\right) + u_i,\ u_i \sim N\left(0, \sigma_u^2\right),\ i = 1, \ldots, n, \tag{2.1}$$

where $y_i$ is the response variable of interest, $\beta_0$ is an overall intercept term, $f_1(x_{1i}), \ldots, f_p(x_{Hi})$ are smooth functions representing the potentially nonlinear effects of $H$ deterministic covariates and $u_i$ are independent and identically distributed normal error terms with variance $\sigma_u^2$.[2] To approximate the potentially nonlinear effects $f_h$, we use the weighted sum of $d_h$ B-spline basis functions, $B_{h1}, \ldots, B_{hd_h}$, such that

$$f_h(x_{hi}) \approx \sum_{j=1}^{d_h} B_{hj}(x_{hi})\beta_{hj} = \mathbf{z}_h^T(x_{hi})\boldsymbol{\beta}_h, \tag{2.2}$$

where $\boldsymbol{\beta}_h$ is a $d_h$-dimensonal column vector of basis coefficients and $\mathbf{z}_h(x_{hi})$ is the $d_h$-dimensonal column vector containing the evaluations of the basis functions at the observed covariate value $x_{hi}$. Thereby, the amount of basis functions and coefficients $d_h$ is steered by the number of knots which divide the domain of the covariate. The bias introduced by the spline representation of a smooth function converges to zero with growing number of knots, see Claeskens et al. (2009) for details. We assume this bias to be negligible by using sufficiently many knots and subsequently postulate equality between an arbitrary smooth function and its spline representation, which leads to the expression

$$f_h(\mathbf{x}_h) = \mathbf{Z}_h\boldsymbol{\beta}_h$$

in compact matrix notation, where

$$\mathbf{Z}_h = \begin{pmatrix} B_{h1}(x_{h1}) & \ldots & B_{hd_h}(x_{h1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hn}) & \ldots & B_{hd_h}(x_{hn}) \end{pmatrix}$$

is a design matrix of dimension $n \times d_h$ assumed to be of full rank. In order to avoid an overfit to the data, a matrix $\mathbf{K}_h$ penalizing to much variability of adjacent coefficients in the coefficient vector $\boldsymbol{\beta}_h$ is assigned to each smooth function resulting in the penalized least squares criterion

$$\left(\mathbf{y} - \beta_0\mathbf{1}_n - \sum_{h=1}^{H}\mathbf{Z}_h\boldsymbol{\beta}_h\right)^T\left(\mathbf{y} - \beta_0\mathbf{1}_n - \sum_{h=1}^{H}\mathbf{Z}_h\boldsymbol{\beta}_h\right) + \sum_{h=1}^{H}\lambda_h\boldsymbol{\beta}_h^T\mathbf{K}_h\boldsymbol{\beta}_h, \tag{2.3}$$

where $\mathbf{y}$ denotes the $n$-dimensional response vector, $\mathbf{1}_n$ is an $n$-dimensonal column vector of ones and $\lambda_h$ is a smoothing parameter determining the impact of the penalty on the minimization criterion. The $d_h \times d_h$-dimensional matrix $\mathbf{K}_h$ of first-order differences, that is penalizing differences of directly contiguous

[2] For notational simplicity, we refrain from adding stochastic covariates and covariates with strictly parametric effects. However, as can be seen in Section 2.5, semiparametric partially linear models can also be handled easily within our framework.

coefficients, has the form

$$\boldsymbol{K}_h = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

Difference matrices of higher orders can be easily constructed. Details on such penalties for B-spline functions can be found, for example, in Eilers and Marx (1996).

Let now $x_h$ be an arbitrary value on the domain of $\mathbf{x}_h$. Defining the smoothing matrix $\mathbf{L}_h(x_h)$ as

$$\mathbf{L}_h(x_h) = (\mathbf{I}_n - \mathbf{S}_{-h})\,\mathbf{Z}_h \left[\mathbf{Z}_h^T(\mathbf{I}_n - \mathbf{S}_{-h})\mathbf{Z}_h + \lambda_h \mathbf{K}_h\right]^{-1} \mathbf{z}_h^T(x_h) \tag{2.4}$$

with $\mathbf{I}_n$ denoting the identity matrix of dimension $n \times n$, $\mathbf{S}_{-h} = \mathbf{Z}_{-h}(\mathbf{Z}_{-h}^T\mathbf{Z}_{-h} + \boldsymbol{\lambda}_{-h}\mathbf{K}_{-h})^{-1}\mathbf{Z}_{-h}^T$ with $\boldsymbol{\lambda}_{-h} = (\lambda_1, \ldots, \lambda_{h-1}, \lambda_{h+1}, \ldots, \lambda_H)$, $\mathbf{Z}_{-h} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{h-1}, \mathbf{Z}_{h+1}, \ldots, \mathbf{Z}_H)$, $\mathbf{K}_{-h} = (\mathbf{K}_1, \ldots, \mathbf{K}_{h-1}, \mathbf{K}_{h+1}, \ldots, \mathbf{K}_H)$, and $\mathbf{z}_h(x_h)$ defined as in (2.2), the estimator of each $f_h(x_h)$ can be written as

$$\hat{f}_h(x_h) = \mathbf{L}_h^T(x_h)\mathbf{y}.$$

It follows that

$$\text{Var}\left[\hat{f}_h(x_h)\right] = \text{Var}\left[\mathbf{L}_h^T(x_h)\mathbf{y}\right] = \mathbf{L}_h^T(x_h)\text{Var}(\mathbf{y})\mathbf{L}_h(x_h) = \mathbf{L}_h^T(x_h)\sigma_u^2\mathbf{I}_n\mathbf{L}_h(x_h) \tag{2.5}$$

holds for homoscedastic and independent errors.

Naturally, the smoothing parameters $\lambda_h$ are unknown. One way to estimate them is to exploit the mixed model representation of penalized splines. In particular, it is always possible to rewrite

$$\mathbf{Z}_h\boldsymbol{\beta}_h = \mathbf{Z}_h(\mathbf{F}_{hf}\boldsymbol{\alpha}_{hf} + \mathbf{F}_{hr}\boldsymbol{\alpha}_{hr}) = \mathbf{X}_{hf}\boldsymbol{\alpha}_{hf} + \mathbf{X}_{hr}\boldsymbol{\alpha}_{hr},$$

where $(\mathbf{F}_{hf}, \mathbf{F}_{hr})$ is of full rank, $\mathbf{F}_{hf}^T\mathbf{F}_{hr} = \mathbf{F}_{hr}^T\mathbf{F}_{hf} = \mathbf{F}_{hf}^T\mathbf{K}_h\mathbf{F}_{hf} = 0$ and $\mathbf{F}_{hu}^T\mathbf{K}_h\mathbf{F}_{hu} = \mathbf{I}_{d_h-q}$ with the difference penalty order $q$.[3] It follows that $X_{hf}$ is of dimension $n \times q$ and $X_{hr}$ is of dimension $n \times (d_h - q)$. Then, $\boldsymbol{\alpha}_{hf}$ contains $q$ fixed coefficients and $\boldsymbol{\alpha}_{hr}$ is a vector of $(d_h - q)$ virtually penalized random coefficients which are assumed to be mutually independent and normally distributed with constant variance $\sigma_{hr}^2$ and independent from the errors $u_i$. In this mixed model formulation, we obtain estimates both for the coefficients (fixed and random) and smoothing parameters by a single (restricted) maximum likelihood estimation. The smoothing parameter estimators $\hat{\lambda}_h = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_{hr}^2}$ are then given as ratios of two (estimated) variances. For details we refer the reader to Ruppert and Wand (2003) or Fahrmeir et al. (2013). In Section 2.3 we will make use of the the mixed model formulation to construct confidence bands.

Asymptotic properties of the penalized spline estimator have been studied, among others, by Claeskens et al. (2009), Kauermann et al. (2009), Wang et al. (2011), Antoniadis et al. (2012), Yoshida and Naito (2012) and Yoshida and Naito (2014). Under mild conditions, consistency of the estimator is shown by Claeskens et al. (2009) for a univariate model with i.i.d. errors. Antoniadis et al. (2012), Yoshida and Naito (2012) and Yoshida and Naito (2014) discuss the asymptotic properties for additive models and derive consistency within different frameworks, always including the case of i.i.d. errors. As we will show, our models can be transformed in such a way that they fit into the class of additive models with i.i.d. errors.

[3] One way to obtain such a decomposition is described in Wood (2006, pp. 316-317).

It should be noted that each row in the initial design matrix $\mathbf{Z}_h$ (i.e., before applying the mixed model reformulation) for each covariate sums up to one, that is, $\sum_{j=1}^{d} B_{hj}(x_{hi}) = 1 \; \forall \, i = 1, \ldots, n$. Obviously, this leads to an identification problem in an additive model with an intercept or multiple smooth components. This issue can be solved by imposing a centering constraint on each function $f_h$ such that

$$\sum_{i=1}^{n} f_h(x_{hi}) = \sum_{i=1}^{n} \mathbf{z}_h^T(x_{hi})\boldsymbol{\beta}_h = 0$$

holds for all $h = 1, \ldots, H$. Following the ideas of Wood (2006, pp. 167-168), this can be achieved by constructing appropriate matrices $\mathbf{W}_h$ of dimension $d_h \times (d_h - 1)$ with orthogonal columns, leading to a reparameterized model with design matrices $\tilde{\mathbf{Z}}_h = \mathbf{Z}_h\mathbf{W}_h$ and penalty matrices $\tilde{\mathbf{K}}_h = \mathbf{W}_h^T\mathbf{K}_h\mathbf{W}_h$. If the mixed model framework is used to determine the smoothing parameters as described above, the reparameterizing procedure to ensure identifiability is done before the mixed model reformulation of the model.

### 2.2.2 Penalized splines for panel data: A first-difference estimator

In comparison to cross-sectional data leading to model (2.1) introduced in the previous section, we now consider individuals (e.g., persons) observed at $T$ consecutive points of time.[4] We therefore consider an additive panel data model

$$y_{it} = \gamma_i + \sum_{h=1}^{H} f_h(x_{hit}) + u_{it}, \; i = 1, \ldots, N, \; t = 1, \ldots, T, \tag{2.6}$$

where $u_{it}$ are assumed to be independent and normally distributed errors with constant variance and $\gamma_i$ are individual-specific, time-invariant fixed effects allowed to be correlated with other covariates. As model (2.6) holds for each point of time, we obtain

$$y_{i,t-1} = \gamma_i + \sum_{h=1}^{H} f_h(x_{hi,t-1}) + u_{i,t-1} \tag{2.7}$$

for a one period time lag. To cancel out the individual-specific effects $\gamma_i$, we subtract (2.7) from (2.6) and obtain

$$\begin{aligned}
\Delta y_{it} = y_{it} - y_{i,t-1} &= \gamma_i - \gamma_i + \sum_{h=1}^{H} \left[ f_h(x_{hit}) - f_h(x_{hi,t-1}) \right] + u_{it} - u_{i,t-1} \\
&= \sum_{h=1}^{H} \left[ \sum_{j=1}^{d_h} B_{hj}(x_{hit})\beta_{hj} - \sum_{j=1}^{d_h} B_{hj}(x_{hi,t-1})\beta_{hj} \right] + \Delta u_{it} \\
&= \sum_{h=1}^{H} \left[ \mathbf{z}_h(x_{hit}) - \mathbf{z}_h(x_{hi,t-1}) \right]^T \boldsymbol{\beta}_h + \Delta u_{it} \\
&= \sum_{h=1}^{H} \left[ \Delta \mathbf{z}_h(x_{hit}) \right]^T \boldsymbol{\beta}_h + \Delta u_{it},
\end{aligned} \tag{2.8}$$

where equation (2.2) is used for the second and third equality and $\Delta$ denotes the first-difference operator over time. Note that only $T - 1$ observations per individual are retained after differencing. Accordingly,

[4] The only reason to refrain from incorporating different observation horizons between persons is notational convenience. As can be seen in Section 2.4 and Section 2.5, unbalanced panels can be handled without any difficulties in our framework.

as the $NT \times d_h$-dimensional design matrix $\mathbf{Z}_h$ of the evaluated basis functions is now given by

$$\mathbf{Z}_h = \begin{pmatrix} B_{h1}(x_{h11}) & \dots & B_{hd_h}(x_{h11}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{h1T}) & \dots & B_{hd_h}(x_{h1T}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN1}) & \dots & B_{hd_h}(x_{hN1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hNT}) & \dots & B_{hd_h}(x_{hNT}) \end{pmatrix}, \tag{2.9}$$

we obtain

$$\Delta \mathbf{y} = \sum_{h=1}^{H} \Delta \mathbf{Z}_h \boldsymbol{\beta}_h + \Delta \mathbf{u} \tag{2.10}$$

in compact matrix notation, where $\Delta \mathbf{y} = (y_{12} - y_{11}, \dots, y_{1T} - y_{1,T-1}, \dots, y_{N2} - y_{N1}, \dots, y_{NT} - y_{N,T-1})^T$ is a $N(T-1)$-dimensional column vector, $\Delta \mathbf{u}$ is defined analogously and the $N(T-1) \times d_h$-dimensional matrix $\Delta \mathbf{Z}_h$ is obtained by building the difference between matrix $\mathbf{Z}_h$ in (2.9) and its one period lagged counterpart:

$$\Delta \mathbf{Z}_h = \begin{pmatrix} B_{h1}(x_{h12}) & \dots & B_{hd_h}(x_{h12}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{h1T}) & \dots & B_{hd_h}(x_{h1T}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN2}) & \dots & B_{hd_h}(x_{hN2}) \\ & \ddots & \\ B_{h1}(x_{hNT}) & \dots & B_{hd_h}(x_{hNT}) \end{pmatrix} - \begin{pmatrix} B_{h1}(x_{h11}) & \dots & B_{hd_h}(x_{h11}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{h1,T-1}) & \dots & B_{hd_h}(x_{h1,T-1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN1}) & \dots & B_{hd_h}(x_{hN1}) \\ \vdots & \ddots & \vdots \\ B_{h1}(x_{hN,T-1}) & \dots & B_{hd_h}(x_{hN,T-1}) \end{pmatrix}.$$

Additionally taking into account penalization, a first-difference penalized spline estimator for all $\boldsymbol{\beta}_h$ can be obtained by minimizing the penalized least squares criterion

$$\left[\Delta \mathbf{y} - \sum_{h=1}^{H} (\Delta \mathbf{Z}_h)\boldsymbol{\beta}_h\right]^T \left[\Delta \mathbf{y} - \sum_{h=1}^{H} (\Delta \mathbf{Z}_h)\boldsymbol{\beta}_h\right] + \sum_{h=1}^{H} \lambda_h {\boldsymbol{\beta}_h}^T \mathbf{K}_h \boldsymbol{\beta}_h. \tag{2.11}$$

Since the smoothing parameters are unknown, one can again exploit the mixed model representation and using (restricted) maximum likelihood estimation as discussed in the previous subsection. Note that the framework is similar to the cross-sectional data case since only the vector of the dependent variable and the design matrices differ between the equations (2.3) and (2.11). The major difference in comparison to cross-sectional data is the problem of autocorrelated errors which are often encountered in panel data contexts. Krivobokova and Kauermann (2007) show that the restricted maximum likelihood based estimation of a smoothing parameter is robust to modest forms of autocorrelation. Moreover, further adjustments on the design matrices and the dependent variable for addressing serial correlation are also possible, see Section 2.3 for an elaboration of this topic.

We briefly have to refer to the identification problem in case of multiple smooth model components: Our aim is to estimate the functions $f_h$, $h = 1, \dots, H$. Hence, model (2.6) should be identified such that

$$\sum_{i=1}^{N} \sum_{t=1}^{T} f_h(x_{hit}) \approx \mathbf{Z}_h \boldsymbol{\beta}_h = 0$$

holds for all $h = 1, \ldots, H$. Therefore, we rewrite the design matrices of the evaluated basis function given in (2.9) and the penalty matrices such that $\tilde{\mathbf{Z}}_h = \mathbf{Z}_h \mathbf{W}_h$ and $\tilde{\mathbf{K}}_h = \mathbf{W}_h^T \mathbf{K}_h \mathbf{W}_h$, proceeding as described in the previous subsection. Furthermore, the identification restriction also implies that a one period lagged design matrix is then constructed directly from $\tilde{\mathbf{Z}}_h$ by taking its one-period-lagged rows. After building the difference between each $\tilde{\mathbf{Z}}_h$ and its respective lagged counterpart, the resulting matrices $\Delta\tilde{\mathbf{Z}}_h$ and the penalty matrices $\tilde{\mathbf{K}}_h$ are plugged into (2.11) to obtain estimators for $\boldsymbol{\beta}_h$ and thus for $f_h$.

Another common approach in fixed effects panel data models is time-demeaning, that is, removing the individual-specific effects $\gamma_i$ by building the mean over time for each individual in equation (2.6) and subtracting the resulting equation from (2.6). Using the information above, this variant is straightforward to derive.

## 2.3 Simultaneous confidence bands for penalized splines

In linear regression models, one is typically interested in the uncertainty of the parameter estimates. Confidence intervals are an established tool to make inferential statements. Similarly, inference about entire smooth functions in nonparametric regression models can be obtained by constructing simultaneous confidence bands around the estimated functions:

$$\left\{ \hat{f}_h(x_h) - c_{h,1-\alpha} \sqrt{\text{Var}\left[\hat{f}_h(x_h)\right]}, \hat{f}_h(x_h) + c_{h,1-\alpha} \sqrt{\text{Var}\left[\hat{f}_h(x_h)\right]}, \, x_{h,min} \leq x_h \leq x_{h,max} \right\}. \tag{2.12}$$

The critical value $c_{h,1-\alpha}$ should ensure that the resulting bands (depending on the sample at hand) cover the true function with a prespecified probability $1 - \alpha$ in all possible samples, that is, $c_{1-\alpha}$ is the $(1 - \alpha)$-quantile of the random variable

$$\sup_{x_{h,min} \leq x_h \leq x_{h,max}} \frac{|\hat{f}_h(x_h) - f_h(x_h)|}{\sqrt{\text{Var}\left[\hat{f}_h(x_h)\right]}}.$$

The difficulty in the penalized spline framework lies in finding the distribution of this random variable. Due to the introduction of a penalty, the estimators for $f_h$, for instance obtained by minimizing (2.3) or (2.11), are usually not unbiased. Krivobokova et al. (2010) propose a solution that takes this bias into account when constructing the simultaneous confidence bands for penalized splines. They consider univariate models while Wiesenfarth et al. (2012) extend the approach to the multivariate case, also covering heteroscedastic errors and spatially heterogeneous splines. The approach performs very well in simulation studies and offers a fast way of inference without the need for computationally intensive resampling procedures. Moreover, the theoretical results from Krivobokova et al. (2010) are appealing as the confidence level does not require a growing sample size to hold, while the average area between the bands decreases with an increasing sample size. The basic idea (derived for the cross-sectional case here) is to exploit the mixed model representation of penalized splines as described in Section 2.2, that is, we consider smooth functions as mixed models:

$$f_h^m(\mathbf{x}_h) := \mathbf{X}_{hf} \boldsymbol{\alpha}_{hf} + \mathbf{X}_{hr} \boldsymbol{\alpha}_{hr} = \mathbf{Z}_h^m \boldsymbol{\beta}_h^m.$$

Recall that both the the random coefficients in each random coefficients vector $\boldsymbol{\alpha}_{hr}$, $h = 1 \dots H$, and the model errors $u_i$ are assumed to be independent and normally distributed with zero expectation and constant variance. Additionally assuming mutual independence, the marginal distribution of $\mathbf{y}$ is given by

$$\mathbf{y} \sim N\left(\beta_0 \mathbf{1}_n + \sum_{h=1}^{H} \mathbf{X}_{hf}\boldsymbol{\alpha}_{hf}, \sigma_u^2 \mathbf{I}_n + \sum_{h=1}^{H} \sigma_{re}^2 \mathbf{X}_{hr}\mathbf{X}_{hr}^T\right).$$

Having such a distribution in a linear mixed model framework, maximum likelihood-type estimation yields the (estimated) best linear unbiased predictor

$$\hat{f}_h^m(x_h) = \mathbf{L}_{m,h}^T(x_h)\mathbf{y} = \mathbf{z}_h^m(x_h)\hat{\boldsymbol{\beta}}_h^m,$$

where $\mathbf{L}_{m,h}(\cdot)$ denotes the smoothing matrix from (2.4) in mixed model formulation with the smoothing parameter $\lambda_h = \frac{\sigma_u^2}{\sigma_{hr}^2}$ assumed to be known and where $\mathbf{z}_h(x_h)$ as defined in (2.2) but now with the basis functions in mixed model formulation. As a linear transformation of the normally distributed variable $\mathbf{y}$, $\hat{f}_h^m(x_h)$ is also normally distributed. It follows that

$$G_h(x_h) = \frac{\mathbf{z}_h^m(x_h)(\hat{\boldsymbol{\beta}}_h^m - \boldsymbol{\beta}_h^m)}{\sqrt{\mathbf{z}_h^m(x_h)\mathrm{Cov}(\hat{\boldsymbol{\beta}}_h^m - \boldsymbol{\beta}_h^m)[\mathbf{z}_h^m(x_h)]^T}} \sim N(0,1) \tag{2.13}$$

is a zero mean Gaussian process, where $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_h^m - \boldsymbol{\beta}_h^m) = \sigma_u^2 \left[\mathbf{Z}_h^T(\mathbf{I}_n - \mathbf{S}_{-h})\mathbf{Z}_h + \lambda_h \mathbf{K}_h\right]^{-1}$ and

$$\mathrm{Cov}\left[G_h(x_{h1}), G_h(x_{h2})\right] = \left[\frac{\mathbf{l}_{m,h}(x_{h1})}{||\mathbf{l}_{m,h}(x_{h1})||}\right]^T \left[\frac{\mathbf{l}_{m,h}(x_{h2})}{||\mathbf{l}_{m,h}(x_{h2})||}\right] =: \boldsymbol{\eta}_{m,h}^T(x_{h1})\boldsymbol{\eta}_{m,h}(x_{h2})$$

with $\mathbf{l}_{m,h}(x_h) = \left[\mathbf{Z}_h^T(\mathbf{I}_n - \mathbf{S}_{-h})\mathbf{Z}_h + \lambda_h \mathbf{K}_h\right]^{-0.5}\mathbf{Z}_h^T(x_h)$.

In practice, the smoothing parameters are estimated from the data such that $\hat{\lambda}_h = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_{hr}^2}$ and thus follow their own distributions with their inherent variabilities. Wood et al. (2016) determine the uncertainty of the smoothing parameter estimator for a generic model class including penalized spline estimators as treated here. We do not consider their more exact calculations and rather follow the arguments of Krivobokova et al. (2010) who prove the smoothing parameter variability to become negligible for growing sample size with the speed of attenuation accelerated by a smaller penalty order. We therefore use second-order penalties in our simulation studies in Section 2.4 and consider equation (2.13) to hold approximately when the smoothing parameters have to be estimated from the data. The advantage from doing so is that Sun and Loader (1994) show that the tail probability of maxima of such zero mean Gaussian processes is determined by

$$\begin{aligned}\alpha &= P\left(\sup_{x_{h,min} \leq x_h \leq x_{h,max}} |G_j(x)| \geqslant c_{h,1-\alpha}\right) \\ &= \frac{\kappa_{m,h}}{\pi}\exp\left(\frac{-c_{h,1-\alpha}}{2}\right) + 2\left[1 - \Phi(c_{h,1-\alpha})\right] + o\left[\exp\left(\frac{-c_{h,1-\alpha}}{2}\right)\right],\end{aligned} \tag{2.14}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and

$$\kappa_{m,h} = \int_{x_{h,min}}^{x_{h,max}} ||\frac{d}{dx}\boldsymbol{\eta}_{m,h}(x)||dx$$

is the length of the mixed model manifold implicitly including the amount of bias which has to be corrected for. Thus, the critical value $c_{h,1-\alpha}$ in (2.12) can be approximately obtained from (2.14). In practice, the error variance $\sigma_u^2$ is estimated from the data and $c_{h,1-\alpha}$ can be calculated from

$$\alpha \approx \frac{\kappa_{m,h}}{\pi}(1 + c_{h,1-\alpha})^{-\zeta/2} + P(|t_\zeta| > c_{h,1-\alpha}),$$

where the random variable $t_\zeta$ follows a t-distribution with $\zeta$ degrees of freedom. Additionally adjusting (2.5) according to the mixed model presentation, that is $\text{Var}\left[\hat{f}_h^m(x_h)\right] = \mathbf{L}_{m,h}^T(x_h)\sigma_u^2\mathbf{I}_n\mathbf{L}_{m,h}(x_h)$, simultaneous confidence bands can then be constructed as

$$\left\{\hat{f}_h^m(x_h) - c_{h,1-\alpha}\sqrt{\text{Var}\left[\hat{f}_h^m(x_h)\right]}, \hat{f}_h^m(x_h) + c_{h,1-\alpha}\sqrt{\text{Var}\left[\hat{f}_h^m(x_h)\right]}, x_{h,min} \leq x_h \leq x_{h,max}\right\}.$$

For further details on such simultaneous confidence bands see Krivobokova et al. (2010) and Wiesenfarth et al. (2012). Their approach is designed for the cross-sectional case, but we will now show how it translates to the panel data context with fixed effects as described in (2.6). Note that the simple, albeit crucial new aspect to contemplate is the serial correlation in the error term $\Delta u_{it}$ of each individual after applying the first-difference transformation described in (2.8). Assuming the $u_{it}$ to be serially uncorrelated, $\Delta u_{it}$ and $\Delta u_{i,t-1}$ exhibit a negative autocorrelation for each individual. In case of a homoscedastic variance, this serial correlation for two consecutive points of time amounts to -0.5, see the appendix for a derivation. We therefore adopt the generalized least squares (GLS) approach and premultiply the differenced model matrix ($\Delta\mathbf{Z}_h$) and the differenced dependent variable $\Delta\mathbf{y}$ in equation (2.11) by $\boldsymbol{\Psi}$, where

$$\boldsymbol{\Psi}\boldsymbol{\Psi}' = \boldsymbol{\Omega}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}_1^{-1} & 0 & \dots & 0 \\ 0 & \boldsymbol{\Omega}_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{\Omega}_N^{-1} \end{pmatrix} \tag{2.15}$$

is a block diagonal matrix with main diagonal block square matrices

$$\boldsymbol{\Omega}_i^{-1} = \begin{pmatrix} 1 & -0.5 & 0 & \dots & 0 \\ -0.5 & 1 & 0 & \dots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & 0 & 1 & -0.5 \\ 0 & 0 & 0 & -0.5 & 1 \end{pmatrix}$$

of dimension $(T-1) \times (T-1)$.[5] Note that, when using first differences and GLS, the smoothing matrix in (2.4) and thus the variance and the confidence bands of the estimated spline curve change accordingly. Provided that there is no heteroscedasticity, applying the GLS transformation on the respective quantities in the penalized least squares criterion (2.11) leads to the ordinary penalized spline case with homoscedastic and uncorrelated errors. Thus, the simultaneous confidence bands described above for cross-sectional data with i.i.d. errors can be applied without any further amendments. Moreover, the asymptotic properties for the penalized spline estimators obtained in such a framework, as discussed in Section 2.2, directly carry over to the panel data context treated here. In the first-difference setting, this is irrespective of whether the number of individuals $N$ or the number of observation periods $T$ grows, see, for example, Wooldridge (2002) for an overview over the asymptotic behavior of fixed effects panel

[5] $\boldsymbol{\Psi}$ can be obtained from $\boldsymbol{\Omega}^{-1}$ with the help of the Cholesky factorization and matrix inversion.

data estimators. The only thing to make sure is that there is some within-variation in the explanatory variables over time as the design matrix $\Delta \mathbf{Z}_h$ only contains differences over time.

In practice, panel data often exhibit additional serial correlation. In the rare cases of an exactly known error structure, the matrices in (2.15) can be adjusted. The common case is that the correlation structure in the error term is unknown and only certain assumptions are made, for example, that errors between different individuals are uncorrelated. In such a case, it is recommended to investigate the residuals for all individuals before or after applying the GLS procedure. If the autocorrelation and partial autocorrelation function suggest the occurrence of a certain underlying autoregressive moving average process, the obtained information could be exploited in the subsequent (iterative) estimation of a feasible GLS estimator with estimated (2.15). Again, the premultiplication of the design matrices and the dependent variable by an appropriate matrix $\hat{\mathbf{\Omega}}$ (asymptotically) leads to a model with uncorrelated and homoscedastic errors if $\mathbf{\Omega}$ is consistently estimated. Another option is a maximum likelihood-type estimation of the model including simultaneous estimation of the autoregressive and moving average parameters. This can be done in a mixed model framework which additionally allows for modeling heteroscedasticity, as described in Pinheiro and Bates (2000).

Wiesenfarth et al. (2012) describe the extension how to build simultaneous confidence bands around the derivatives. In the case of B-spline basis functions, the derivative of the smoothing matrix in (2.4) for the cross-sectional case is given by

$$\mathbf{L}_h'(x_h) = (\mathbf{I}_n - \mathbf{S}_{-h})\, \mathbf{Z}_h \left[\mathbf{Z}_h^T (\mathbf{I}_n - \mathbf{S}_{-h})\, \mathbf{Z}_h + \lambda_h \mathbf{Z}_h\right]^{-1} \left[\mathbf{z}_h'(x_h)\right]^T,$$

where $\left[\mathbf{z}_h'(x_h)\right]^T$ denotes the row vector of the derivatives of the initial basis functions, evaluated at some value $x_h$ (see De Boor, 2001, Ch. 10). Thus, derivative estimates are practically obtained with negligible effort once a penalized least squares criterion like (2.3) has been minimized. Critical values and simultaneous confidence bands for the derivatives, also for panel data settings, can then be obtained by analogy with the steps described above.

## 2.4 Simulation studies

We consider data generated from model (2.6) with the individual-specific fixed effects $\gamma_i = i$ and the $H = 3$ true functions

$$\begin{aligned} f_1(x_{1it}) &= \sin^2\left[2\pi(x_{1it} - 0.5)\right], \\ f_2(x_{2it}) &= 0.6 b_{30,17}(x_{2it}) + 0.4 b_{3,11}(x_{2it}), \\ f_3(x_{3it}) &= x_{3it}(1 - x_{3it}), \end{aligned}$$

with $b_{l,m}(x) = \Gamma(l+m)\left[\Gamma(l)\Gamma(m)\right]^{-1} x^{l-1}(1-x)^{m-1}$, where $\Gamma(r)$ denotes the gamma function. All functions were also considered in Wiesenfarth et al. (2012). They are scaled such that their standard deviations are equal to one. The functions and their derivatives are shown in Figure 2.3 in the appendix. The errors are generated as i.i.d. Gaussian errors with standard deviation $\sigma_u = 0.5$. We consider an unbalanced panel data design with total sample sizes of $n = (525, 1050, 2100)$, where $N = (75, 150, 300)$ imaginary individuals are observed over different time horizons without breaks, that is, there are no missing observations between the first and last point of time at which one individual is observed. Note that due to taking first differences according to (2.10), the sample size used for the estimation decreases by the number of individuals, that is, we obtain the effective sample sizes $n - N = (450, 900, 2700)$. The

Table 2.1: Coverage rates in simulations, average areas between confidence bands in parentheses. Columns (i) denote estimation with using GLS, columns (ii) without using GLS.

| $n$ | $f_1$ | | $f_2$ | | $f_3$ | |
|---|---|---|---|---|---|---|
| | (i) | (ii) | (i) | (ii) | (i) | (ii) |
| 525 | 0.95 | 0.86 | 0.93 | 0.85 | 0.97 | 0.91 |
| | (3.42) | (3.48) | (3.67) | (3.63) | (3.07) | (3.24) |
| 1050 | 0.95 | 0.88 | 0.95 | 0.85 | 0.97 | 0.88 |
| | (2.45) | (2.51) | (2.45) | (2.44) | (2.12) | (2.14) |
| 2100 | 0.95 | 0.84 | 0.96 | 0.88 | 0.97 | 0.88 |
| | (1.81) | (1.80) | (1.90) | (1.88) | (1.55) | (1.55) |
| | $f_1'$ | | $f_2'$ | | $f_3'$ | |
| | (i) | (ii) | (i) | (ii) | (i) | (ii) |
| 525 | 0.90 | 0.75 | 0.80 | 0.62 | 0.94 | 0.86 |
| | (30.60) | (31.71) | (32.80) | (33.57) | (17.82) | (19.29) |
| 1050 | 0.90 | 0.77 | 0.85 | 0.66 | 0.94 | 0.84 |
| | (23.51) | (24.42) | (25.19) | (25.85) | (14.06) | (14.92) |
| 4200 | 0.85 | 0.70 | 0.73 | 0.60 | 0.95 | 0.83 |
| | (15.23) | (15.69) | (17.01) | (17.26) | (8.77) | (9.19) |

covariates for each individual are taken to be distributed over $\{a - 0.04, a - 0.03, \ldots, a, a + 0.01\}$ with

$$P(X = x) = \begin{cases} 0.5, & \text{if } x = a, \\ 0.1 & \text{else,} \end{cases}$$

with $a$ being randomly drawn with equal probability from $\{0.04, 0.05, ..., 0.99\}$ for each individual. This setting is designed to mimic a real-world panel dataset where covariate values of individuals are often restricted to a finite set of values and can sometimes remain constant over time. In all settings, we take 40 equidistant knots for all covariates. The results are based on 500 Monte Carlo replicates and a nominal coverage rate of 95%. Note that under the error assumptions stated above, the errors after building first differences are serially correlated (see Section 2.3). We use B-spline basis function of degree three and impose a penalty on second-order differences of the B-spline coefficients.

In Table 2.1, the resulting coverage rates with and without using GLS are shown. It can be seen that not taking into account the autocorrelation in the error term leads to substantial undercoverage. In contrast, even for moderate sample size, the confidence bands estimated by GLS generally perform quite accurately, that is, the nominal coverage is met. Likewise, the average area between the bands decreases with growing sample size. These results are in line with those of Wiesenfarth et al. (2012).

Using the same setting, we also examine the coverage rates of the confidence bands for the derivatives. The results in Table 2.1 show adequate coverage rates for the comparably simple linear derivative $f_3'(x_{3it})$ but not for the two other more complicated functions. Especially the confidence bands for $f_2'(x_{2it})$ perform poorly,[6] even if the sample size is huge ($n = 4200$) or the error variance is low (not shown here for brevity). In further simulations, we also varied the number of knots and the difference orders for the penalty. Although sometimes observing improvements in the coverage rates (with or without the expense of wider confidence bands), we did not find a distinct pattern how to reach the nominal coverage rate. Thus, we can only advise to be careful in making inferential statements about the derivatives of potentially sophisticated curves.

In addition, we replicated the simulation studies with non-Gaussian errors and autocorrelated errors. The

[6] We observe similar problems for other functions, for example, $f(x) = (0.5 - x)^3$, see the online supplement.

results are comparable to our simulation setting with independent Gaussian errors. The robustness of the here proposed confidence bands for non-normal symmetric error distributions was also demonstrated by Loader and Sun (1997). Furthermore, our simulations indicate that slight violations of the serial independence assumption are not too harmful. However, as shown above, disregarding major serial correlation as introduced by the first-difference transformation to uncorrelated errors is problematic. Thus, we advise the practitioner to investigate the residuals and apply, if necessary, more adequate modeling approaches as described in section 2.3.

As for all fixed effects panel data models, it is also important to ensure that there is sufficient intra-personal variation for all covariates. If this is not the case, the model matrix after applying first differences contains many zeros and thus, there is too little variation to estimate the function adequately.

The results for additional simulations not shown but discussed in this section can be found in the online supplement.

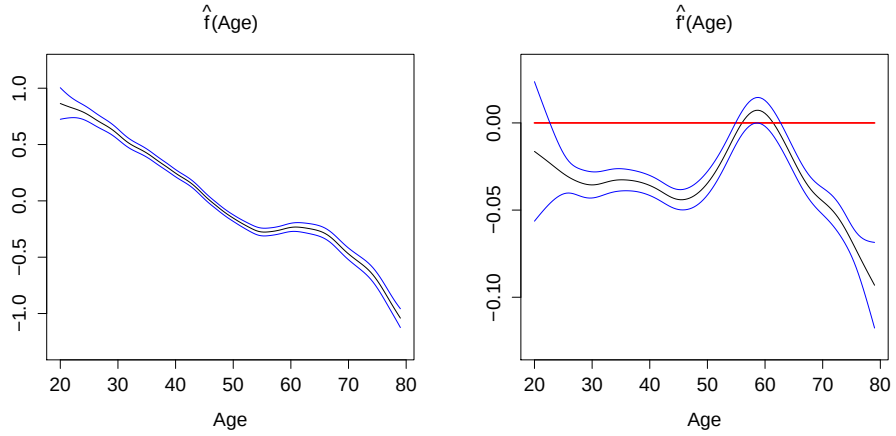
## 2.5 Studying the relationship between aging and life satisfaction

There is a considerable strand of literature studying how life satisfaction evolves over the lifespan. So far, there is no broad consensus on the shape of this relationship, as study results differ while applying different methodologies and datasets. A recent overview on this topic is given by López Ulloa et al. (2013). Frequently, an a priori specified U-shaped relationship is tested in a parametric way. One exception is the work of Wunder et al. (2011), who apply a semiparametric random effects model using the SOEP and the British Household Panel Survey. However, they do not address possible endogeneity of time-invariant omitted covariates which can be done by incorporating individual-specific time constant fixed effects. In the context of the relationship between aging and life satisfaction, the importance of doing so is highlighted by Ferrer-i Carbonell and Frijters (2004). Using fixed effects panel models, Frijters and Beatton (2012) apply a quite flexible step function based on 5-year-intervals for the influence of age on life satisfaction, which is, however, non-continuous and does not allow for uncertainty statements. To the best of our knowledge, we provide the first fully flexible fixed effects panel data approach also allowing for statistical quantification of uncertainty. To illustrate our method, we use SOEP data from 1994 to 2011, see Wagner et al. (2007) for details on the dataset. Following the results of Ferrer-i Carbonell and Frijters (2004), we treat the life satisfaction score[7], which is measured on an actually ordinal 11-point scale ranging from 0 (completely dissatisfied) to 10 (completely satisfied), as cardinal. While applying a first-difference estimator, the effects on life satisfaction are assumed to be exclusively instantaneous, that is, an increase or decrease of an explanatory variable in one year influences life satisfaction solely in the same year. This is questionable especially in the case of certain life events like changes in the marital status, for instance. Therefore, we follow an approach similar to Laporte and Windmeijer (2005) and add dummy variables for each of the two years before and after a life event,[8] including changes in marital, employment and disability status. Furthermore, we include nonparametric effects for age and net household income (with 60 equidistant knots each) and linear effects for household size and nights

[7] The corresponding question in the SOEP survey ist: "How satisfied are you with your life, all things considered?"

[8] Incorporating leads and lags results in a smaller sample size. In our case, we require an individual to be observed in at least six consecutive periods corresponding to at least one observation for estimation after building two leads and two lags and taking first differences. Albeit the loss of observations, this modeling procedure allows us to investigate whether effects on life satisfaction are long-lasting or just temporary, see, for instance, Lucas (2007) for a discussion on this issue.

Figure 2.1: Estimated nonparametric relationship between age and life satisfaction with confidence bands (left panel), corresponding estimated derivative (right panel)




stayed in hospital in the previous year. Thus, our model to estimate is

$$\text{Life Satisfaction}_{it} = \gamma_i + f(\text{Age}_{it}) + f(\text{Household Income}_{it}) + \mathbf{c}_{it}^T\delta + u_{it}, \tag{2.16}$$

where the vector $\mathbf{c}_{it}$ captures the values of all variables (including lags and leads) which are modeled in a parametric fashion. The final sample size after removing missing values amounts to $n = 143,299$.

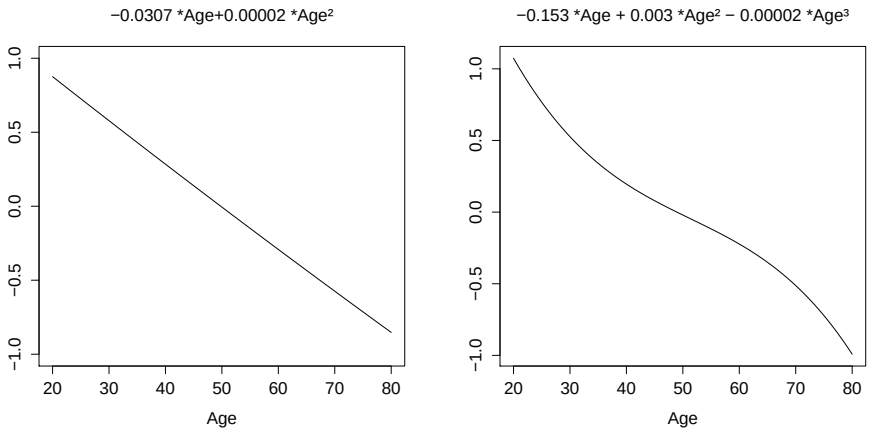
The results for the nonparametric effect of age on life satisfaction can be found in the left panel of Figure 2.1.

It can be seen that young people tend to become more and more unhappy as they become older. This decrease in life satisfaction is stopped and even slightly reversed at the age of around 60 for a couple of years. After that, increasing age again goes along with a reduction in life satisfaction. The estimated derivative of this effect and its confidence bands are shown in the right panel of Figure 2.1.[9] For ages older than about 25 years, the zero line is not covered by the bands over almost the whole life span, indicating a significant negative effect of age on life satisfaction within these ages. This does not hold for the ages around 60 years. There, the confidence bands cover the zero line and the lower band almost crosses the zero line once. With regard to our simulation studies in Section 2.4, however, these results should be taken with caution.

For comparison, we also estimate two simple parametric first-difference panel data models, where the smooth functions of age and household income in equation (2.16) are replaced by quadratic and cubic polynomials. The results for the age effect can be found in Figure 2.2. It can be seen that the quadratic fit is a quasi-linear decreasing function, whereas the cubic fit shows some curvature while still exhibiting a clear downward trend over the lifespan. Neither of these estimated functions can capture the stage of constant or even increasing life satisfaction for the ages around 60 years. Thus, it is advisable to use a nonparametric estimator here to estimate the relationship of interest. In our analysis the often found U-shape or any other simple relationship between age and life satisfaction cannot be confirmed. Qualitatively, our results rather resemble those of Wunder et al. (2011). The nonparametric effect of net household income as well as the purely parametric effects are shown in Figure 2.4 and Table 2.2 in the appendix.

[9] In fact, the estimated derivative was obtained in a new estimation with spline degree five and third-order difference penalty, leading to a smoother curve.

Figure 2.2: Estimated relationship between age and life satisfaction based on a squared (left) and a cubic polynomial (right)




## 2.6 Discussion and conclusions

In this paper, we enhanced the statistical toolbox by presenting a nonparametric first-difference estimator for fixed effects panel data models based on penalized splines combined with a corresponding fast way of inference via simultaneous confidence bands. Our approach allows to estimate and draw inferences from fixed effects panel data models in a highly flexible way and without a priori specifications of covariate effects. One further merit of our methodology is that numerous covariates, either modeled in a parametric or nonparametric way, can be handled easily. Moreover, the derivatives of the estimated effects as well as of their confidence bands can be obtained with negligible additional effort. The proposed approach is available for practitioners in the new `R` package `pamfe` which enables the fast estimation of nonparametric and semiparametric partially linear models even for large sample sizes. Using data from the SOEP, we illustrated the applicability our method by investigating the relationship between age and life satisfaction. We found that it is not advisable to model this nonlinear relationship in a strictly parametric fashion. Simulation studies showed an overall good performance of our method with the exception of the confidence bands for the derivatives which sometimes failed to hit the nominal coverage rate. A possible explanation is that the smoothing parameters are estimated and optimized for the original functions and not for the derivatives, as pointed out by Ruppert and Wand (2003, Ch. 6.8). It might be an interesting direction for future research to address this problem, maybe in a fully Bayesian framework. However, aside from the higher computational effort generally required by Bayesian methods, Bayesian credible bands tend to be conservative from a frequentist point of view, as shown by Krivobokova et al. (2010).

# 2.A Appendix

### Serial correlation in the first-difference errors

Consider equation (2.6): If the error terms $u_{it},\ i = 1,\ldots,N,\ t = 1,\ldots,T_i$ are homoscedastic and independent with expectation zero, then $\mathrm{E}(u_{it}u_{i,t-1}) = 0$ and $\mathrm{E}(u_{it}u_{it}) = \sigma_u^2$. It follows for the errors $\triangle u_{it} = u_{it} - u_{i,t-1}$ in equation (2.8):

$$\mathrm{E}(\Delta u_{it}) = \mathrm{E}(u_{it} - u_{i,t-1}) = 0$$

and

$$\mathrm{Var}(\Delta u_{it}) = \mathrm{Var}(u_{it} - u_{i,t-1}) = \mathrm{Var}(u_{it}) + \mathrm{Var}(u_{i,t-1}) = 2\sigma_u^2.$$

The correlation of two consecutive error terms for the same individual after applying first differences is then given by

$$\begin{aligned}
\mathrm{Cor}(\Delta u_{it}, \Delta u_{i,t-1}) &= \frac{\mathrm{E}\left[(\Delta u_{it})(\Delta u_{i,t-1})\right]}{\sqrt{\mathrm{Var}(\Delta u_{it})\mathrm{Var}(\Delta u_{i,t-1})}} \\
&= \frac{\mathrm{E}\left[(u_{it} - u_{i,t-1})(u_{i,t-1} - u_{i,t-2})\right]}{\sqrt{2\sigma_u^2 2\sigma_u^2}} \\
&= \frac{\mathrm{E}(-u_{i,t-1}^2)}{2\sigma_u^2} = \frac{-\sigma_u^2}{2\sigma_u^2} = -0.5.
\end{aligned}$$

## Figures and tables

Figure 2.3: Simulation studies: True, scaled functions (left) and corresponding derivatives (right).

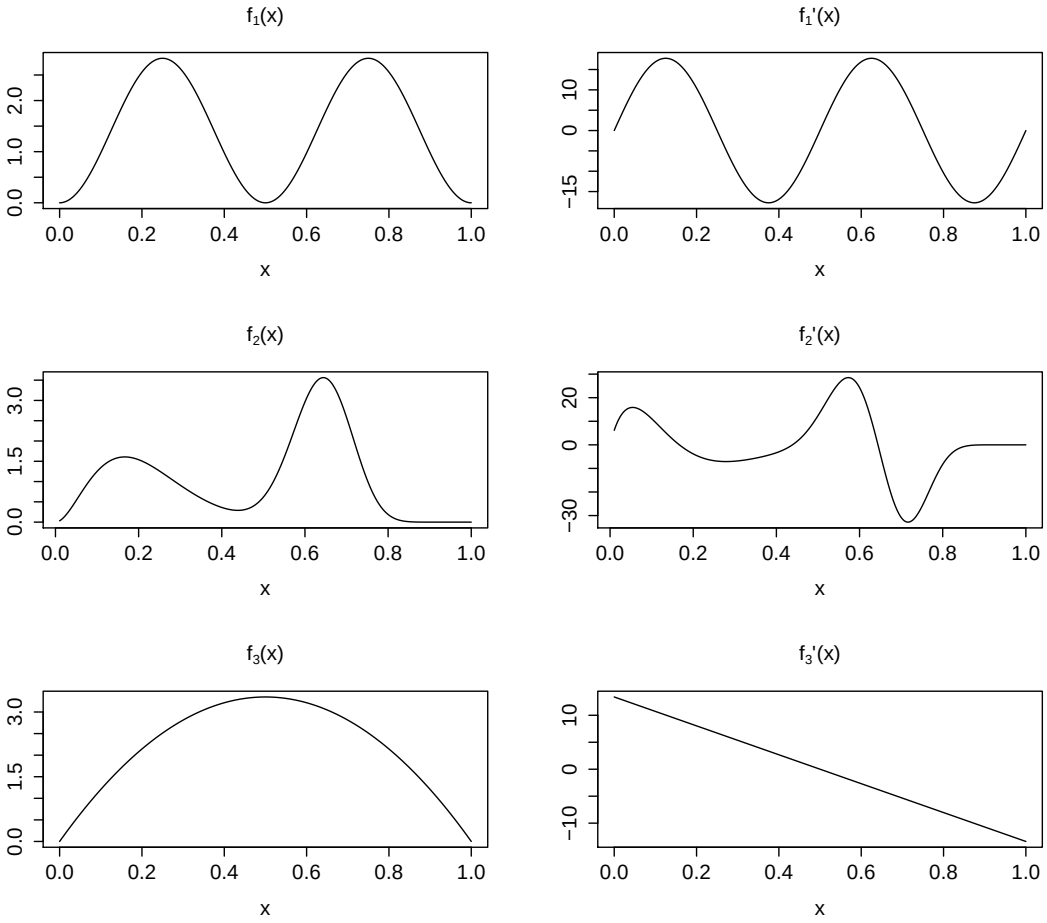



Figure 2.4: Nonparametrically estimated relationship between household income (in 1000 €) and life satisfaction with confidence bands

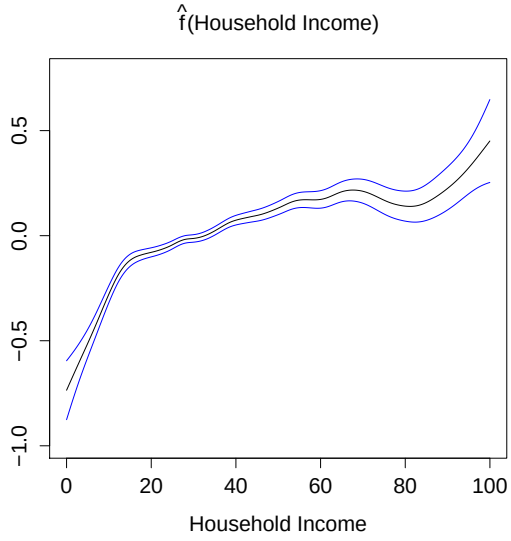


Table 2.2: Estimation results for strictly parametric components. Note that the reference categories for the marital status and its leads and lags are "single" and its respective leads and lags. For the disability status "not disabled" serves at reference category, so does "nonworking" for the employment status.

| Variable | Coefficient | P-value |
|---|---|---|
| Household size | -0.0048 | 0.5668 |
| Nights in hospital | -0.0102 | 0.0000 |
| Disability Status: Disabled + 2 years | -0.0156 | 0.5107 |
| Disability Status: Disabled + 1 year | 0.0334 | 0.1763 |
| Disability Status: Disabled | -0.1533 | 0.0000 |
| Disability Status: Disabled - 1 year | -0.2208 | 0.0000 |
| Disability Status: Disabled - 2 years | -0.1775 | 0.0000 |
| Divorced + 2 years | 0.0482 | 0.2165 |
| Divorced + 1 year | 0.2686 | 0.0000 |
| Divorced | 0.0289 | 0.5528 |
| Divorced - 1 year | -0.1348 | 0.1095 |
| Divorced - 2 years | -0.0744 | 0.3061 |
| Widowed + 2 years | 0.2420 | 0.0000 |
| Widowed + 1 year | 0.5067 | 0.0000 |
| Widowed | -0.3942 | 0.0000 |
| Widowed - 1 year | -0.0820 | 0.2459 |
| Widowed - 2 years | -0.0935 | 0.1195 |
| Married + 2 years | -0.1082 | 0.0006 |
| Married + 1 year | -0.0569 | 0.1388 |
| Married | 0.1143 | 0.0046 |
| Married - 1 year | 0.1463 | 0.0007 |
| Married - 2 years | 0.1418 | 0.0002 |
| Part time employed | 0.0061 | 0.7807 |
| Full time employed | 0.1235 | 0.0000 |
| Unemployed | -0.4843 | 0.0000 |

# 3 Treatment effects beyond the mean using GAMLSS

# Treatment effects beyond the mean using GAMLSS

Maike Hohberg [*], Peter Pütz [†], Thomas Kneib [‡]

**Abstract**

This paper introduces distributional regression, also known as generalized additive models for location, scale and shape (GAMLSS), as a modeling framework for analyzing treatment effects beyond the mean. By relating each parameter of the response distribution to explanatory variables, GAMLSS model the treatment effect on the whole conditional distribution. Additionally, any non-normally distributed outcome and nonlinear effects of explanatory variables can be incorporated. We elaborate on the combination of GAMLSS with program evaluation methods in economics and provide practical guidance on the usage of GAMLSS by reanalyzing data from the Mexican Progresa program. Contrary to expectations, no significant effects of a cash transfer on the conditional inequality level between treatment and control group are found.

**Keywords**: Conditional distribution; GAMLSS; Impact evaluation; Inequality; Treatment effects

[*]Corresponding author: Maike Hohberg, Chair of Statistics, Economics Faculty, University of Goettingen, Humboldtallee 3, 37073 Goettingen, mhohber@uni-goettingen.de.

[†]Chair of Statistics, Economics Faculty, University of Goettingen.

[‡]Chair of Statistics, Economics Faculty, University of Goettingen.

We thank David McKenzie (World Bank), Jörg Langbein (KfW and World Bank), and Marion Krämer (DEval) for comments on an earlier draft.

## 3.1 Introduction

Program evaluation typically identifies the effect of a policy or a program on the mean of the response variable of interest. This effect is estimated as the average difference between treatment and comparison group with respect to the response variable, potentially controlling for confounding covariates. However, questions such as "How does the treatment influence a person's future income distribution" or "How does the treatment affect consumption inequality conditional on covariates" cannot be adequately answered when evaluating mean effects alone. Concentrating on mean differences between a treatment group and a comparison group is likely to miss important information about changes along the whole distribution of an outcome, for example in terms of an unintended increase in inequality, or when targeting ex ante vulnerability to a certain risk. These are economic concepts that do not only take the expected mean into account but rely on other measures such as the variance and skewness of the response.

As shown recently by Bitler et al. (2017), analyzing average effects in subgroups does not adequately capture heterogeneities along the outcome distribution. For a systematic and coherent analysis of treatment effects on all functionals of the response distribution, we introduce generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005) to the evaluation literature. GAMLSS allow all parameters of the response distribution to vary with explanatory variables and can hence be used to assess how the conditional response distribution changes due to the treatment. In addition, GAMLSS constitute an overarching framework to easily incorporate nonlinear, random, and spatial effects. Hence, the relationship between the covariates and the predictors can be modeled very flexibly, for example by using splines for nonlinear effects or Gaussian-Markov random fields for spatial information. The method encompasses a wide range of potential outcome distributions, including discrete and multivariate distributions, and distributions for shares. Due to estimating only *one* model including all distributional parameters, practically every distribution functional (quantiles, Gini coefficient, etc.) can be derived consistently from the conditional distribution making the scope of application manifold.

Besides a brief review of the methodological background for GAMLSS, our main aim is to practically demonstrate how to implement them in the course of treatment effects and what additional information can be drawn from those models. For this, we have chosen an example that is very familiar to the evaluation community: We rely on the same household survey used in Angelucci and De Giorgi (2009) to evaluate Progresa/Oportunidades/Prospera - a cash transfer program in Mexico. Initiated in 1997, the experimental design of the program allocated cash transfers to poor families in treatment villages in exchange for the households' children regularly attending school and for utilizing preventive care measures regarding health and nutrition. By using this extensively researched program as our application example, we show additional results using GAMLSS. In fact, we find no significant decline in food consumption inequality after the introduction of conditional cash transfers - a result that has gone unnoticed in the several analyses of the program's heterogeneous effects (e.g., Djebbari and Smith, 2008; Chavez-Martin del Campo, 2006).

While GAMLSS have not been used in the context of program evaluation, there is a substantial strand of literature that focuses on treatment effects on the whole distribution of an outcome or, to put it differently, on building counterfactual distributions. The idea is to consider the distribution of the treated versus their distribution if they had not been treated. The literature generally differentiates between effects on the unconditional distribution and the conditional distribution. While the effects on the unconditional distribution and unconditional quantile effects have been dealt with in Firpo (2007), Firpo et al. (2009), Rothe (2010), Rothe (2012) and Frölich and Melly (2013), for example, the focus of this paper is the conditional distribution and the functionals that can be derived from it. Conditional distributions are

of interest, when analyzing the effect heterogeneity based on the observed characteristics (Frölich and Melly, 2013). Especially in the case of inequality, conditional distributions are important to differentiate between within and between variance. For example, differences in consumption or income might stem from different characteristics or abilities such as years of education. With conditional distributions, we, however, assess the differences in consumption or income for individuals with equal or similar education and work experience. The fair notion would be that a person with higher education and more work experience earns more. It is the conditional inequality that is perceived as unfair.

To estimate the conditional distribution, a popular approach is to use quantile regression (Koenker and Bassett, 1978; Koenker, 2005). Quantile regression is a very powerful instrument if one is interested in the effect at a specific quantile. However, distributional characteristics can be derived only after the effects at a very high number of quantiles have been estimated which then yields an approximation of the whole distribution. For example, Machado and Mata (2005), Melly (2005), Angrist et al. (2006), and Chernozhukov and Hansen (2006) considered effects over a set of quantiles. The conditional distribution obtained via quantile regression can be integrated over the range of covariates to get the effects on the unconditional distribution. As we believe that quantile regression is most familiar to practitioners when estimating effects beyond the mean, we will elaborate a direct comparison of GAMLSS and quantile regression in Section 3.3.

Other interesting approaches to go beyond the mean in regression modeling include Chernozhukov et al. (2013) and Chernozhukov et al. (2018) who introduce "distribution regression". Building upon Foresi and Peracchi (1995), they develop models that do not assume a parametric distribution but estimate the whole conditional distribution flexibly. The basic idea is to estimate the distribution of the dependent variable via several binary regressions for $F(z|x_i) = Pr(y_i \leq z|x_i)$ based on a fine grid of values $z$. These models have the advantage of not requiring an assumption about the form of the response distribution. However, they require constrained estimates to avoid crossing predictions similar to crossing quantiles in quantile regression. Recently, Shen (2019) proposed a nonparametric approach based on kernel functions to estimate the effect of minimum wages on the conditional income distribution. She points out that the flexibility of estimating distributional effects conditional on the other covariates is also useful for the regression discontinuity design (RDD). In Shen and Zhang (2016) they develop tests relating the stochastic dominance testing to the RDD.

Thus, different concepts are already introduced with different scope for application. By applying GAMLSS to the evaluation context, we provide a flexible, parametric complement to the existing approaches. The advantage of this approach is that it provides *one* coherent model for the conditional distribution which estimates simultaneously the effect on all distributional parameters avoiding crossing quantiles or crossing predictions. If the distributional assumption is appropriate, the parametric approach allows us to rely on classical results for inference in either frequentist or Bayesian formulations, including large sample theory. The parametric formulation furthermore enables us to derive various quantities of interest from the same estimated distribution (quantiles, moments, Gini coefficient, interquartile range, etc.) which are all consistent with each other. As the distributional assumption obviously plays a crucial role in GAMLSS, we suggest guiding steps and easy-to-use tools for the practitioner to decide on a distribution.

The remainder is structured as follows: Section 3.2 provides the methodological background of GAMLSS. Section 3.3 elaborates on the potential benefits and limitations of GAMLSS for evaluating treatment effects. A practical step-by-step implementation and interpretation is given in Section 3.4. Though this section uses data from a randomized controlled trial (RCT), the methodology proposed in this paper applies to non-experimental methods as well. The appendix elaborates on the combination of GAMLSS

with other evaluation methods including panel data approaches, difference-in-differences, instrumental variables (IV), and regression discontinuity design (RDD). Section 3.5 concludes.

## 3.2 Generalized additive models for location, scale and shape

### 3.2.1 A general introduction to GAMLSS

For the sake of illustration, we start with a basic regression as it would be used, for example, when evaluating data from an RCT. Based on observed values $(\mathbf{x}_i', T_i, y_i)$, $i = 1, \ldots, n$, we are interested in determining the regression relation between a treatment, $T_i$, and the response variable $y_i$, while controlling for a vector of non-stochastic covariates $\mathbf{x}_i'$. For simplicity and in line with the application in Section 3.4, we describe the method in the context of a binary treatment but it applies to the continuous case as well. A corresponding simple linear model

$$y_i = \beta_0 + \beta_T T_i + \mathbf{x}_i' \boldsymbol{\beta}_1 + \varepsilon_i$$

with error terms $\varepsilon_i$ subject to $E(\varepsilon_i) = 0$ implies that the treatment and the remaining covariates linearly determine the expectation of the response via

$$E(y_i) = \mu_i = \beta_0 + \beta_T T_i + \mathbf{x}_i' \boldsymbol{\beta}_1.$$

If, in addition, the distribution of the error term is assumed to not functionally depend on the observed explanatory variables (implying, for example, homoscedasticity), the model focuses exclusively on the expected value, that is, it is a mean regression model. In other words, all effects that do not affect the mean but other parameters of the response distribution such as the scale parameter are implicitly subsumed into the error term.

One possibility to weaken the focus on the mean and give more structure to the remaining effects is to relate all parameters of a response distribution to explanatory variables. In the case of a normally distributed response $y_i \sim N(\mu_i, \sigma_i^2)$, both mean and variance could depend on the explanatory variables. Assuming again one treatment variable $T_i$ and additional covariates $\mathbf{x}_i'$, the corresponding relations in a GAMLSS can be specified as follows:

$$\mu_i = \beta_0^\mu + \beta_T^\mu T_i + \mathbf{x}_i' \boldsymbol{\beta}_1^\mu, \tag{3.1}$$

$$\log(\sigma_i) = \beta_0^\sigma + \beta_T^\sigma T_i + \mathbf{x}_i' \boldsymbol{\beta}_1^\sigma. \tag{3.2}$$

Here, the superscripts in $\beta_0^\mu, \beta_T^\mu, \boldsymbol{\beta}_1^\mu, \beta_0^\sigma, \beta_T^\sigma$ and $\boldsymbol{\beta}_1^\sigma$ indicate the dependency of the intercepts and slopes on the respective distribution parameters. The log transformation in (3.2) is applied in order to guarantee positive standard deviations for any value of the explanatory variable.

Aside from the normal distribution, a wide range of possible distributions is incorporated in the flexible GAMLSS framework:

(a) In addition to distributions with location and scale parameters, distributions with skewness and kurtosis parameters can be modeled.

(b) For count data, not only the Poisson but also alternative distributions that account for over-dispersion and zero-inflation can be used.

(c) Often we consider nonnegative dependent variables (e.g., income) with an amount of zeros that cannot be captured by continuous distributions. For these cases, a mixed discrete-continuous distribution can be used that combines a nonnegative continuous distribution with a point mass in zero.

(d) For response variables that are shares (also called fractional responses) we can consider continuous distributions defined on the unit interval.

(e) Even multivariate distributions, that is, where the response is a vector of dependent variables, can be placed within this modeling framework (Klein, Kneib, Klasen and Lang, 2015).

GAMLSS assume that the observed $y_i$ are conditionally independent and that their distribution can be described by a parametric density $p(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK})$ where $\vartheta_{i1}, \ldots, \vartheta_{iK}$ are $K$ different parameters of the distribution. For each of these parameters we can specify an equation of the form

$$g_k(\vartheta_{ik}) = \beta_0^{\vartheta_k} + \beta_T^{\vartheta_k} T_i + x_i' \boldsymbol{\beta}^{\vartheta_k},$$

where the link function $g_k$ ensures the compliance with the requirements of the parameter space (such as the log link to ensure positive variances in Equation (3.1)). Linking the parameters to an unconstrained domain also facilitates the consideration of semiparametric, additive regression specifications including, for example, nonlinear, spatial or random effects. Due to assuming a distribution for the response variable, model estimation can be done by maximum likelihood (Rigby and Stasinopoulos, 2005) or Bayesian methods (Klein, Kneib, Lang and Sohn, 2015).

### 3.2.2 Additive predictors

The univariate case described in the previous subsection can be easily extended to a multivariate and even more flexible setting. In particular, each parameter $\vartheta_{ik}, k = 1, \ldots, K$, of the response distribution is now conditioned on several explanatory variables and can be related to a predictor $\eta_i^{\vartheta_k}$ via a link function $g_k$ such that $\vartheta_{ik} = g_k^{-1}(\eta_i^{\vartheta_k})$.

A generic predictor for parameter $\vartheta_{ik}$ takes on the following form:

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + \beta_T^{\vartheta_k} T_i + f_1^{\vartheta_k}(\mathbf{x}_{1i}) + \cdots + f_{J_k}^{\vartheta_k}(\mathbf{x}_{J_k i}).$$

This representation shows nicely why we refer to $\eta_i^{\vartheta_k}$ as a "structured additive predictor". While $\beta_0^{\vartheta_k}$ denotes the overall level of the predictor and $\beta_T^{\vartheta_k}$ is the effect of a binary treatment on the predictor, functions $f_j^{\vartheta_k}(\mathbf{x}_{ji}), j = 1, \ldots, J_k$, can be chosen to model a range of different effects of a vector of explanatory variables $\mathbf{x}_{ji}$:

(a) Linear effects are captured by linear functions $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = x_{ji}\beta_j^{\vartheta_k}$, where $x_{ji}$ is a scalar and $\beta_j^{\vartheta_k}$ a regression coefficient.

(b) Nonlinear effects can be included for continuous explanatory variables via smooth functions $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = f_j^{\vartheta_k}(x_{ji})$ where $x_{ji}$ is a scalar. We recommend using P(enalized)-splines (Eilers and Marx, 1996) in order to include potentially nonlinear effects of continuous variables.

(c) An underlying spatial pattern can be accounted for by specifying $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = f_j^{\vartheta_k}(s_i)$, where $s_i$ is some type of spatial information such as geographical coordinates or administrative units.

(d) If the data are clustered, random or fixed effects $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = \beta_{j,g_i}^{\vartheta_k}$ can be included with $g_i$ denoting the cluster the observations are grouped into.

Consequently, GAMLSS allows the researcher to incorporate very different types of effects within one modeling framework. Estimation may then be done via a back-fitting approach within the Newton-Raphson type algorithm that maximizes the penalized likelihood and estimates the unknown quantities simultaneously. The methodology is implemented in the `gamlss` package in the software `R`, and described extensively in Stasinopoulos and Rigby (2007) and Stasinopoulos et al. (2017). Alternatively, a Bayesian implementation is available in the open source software `BayesX` (Belitz et al., 2015).

### 3.2.3 GAMLSS vs. quantile regression

A popular alternative to simple mean regression is quantile regression, see, for example, Koenker (2005) for an excellent introduction. Quantile regression relates not the mean but quantiles of the outcome variable to explanatory variables without making a distributional assumption about the outcome variable. In addition to requiring independence of observed values $y_i$, a quantile regression model with one explanatory variable $x_i$ only assumes that

$$y_i = \beta_{0,\tau} + \beta_{1,\tau} x_i + \varepsilon_{i,\tau}$$

where $\varepsilon_{i,\tau}$ is a quantile-specific error term with the quantile condition $P(\varepsilon_{i,\tau} \leq 0) = \tau$ replacing the usual assumption $E(\varepsilon_{i,\tau}) = 0$. This implies a specific form of the relationship: The explanatory variable influences the $\tau$-quantile in a linear fashion. Thus, the model can still be misspecified even though we do not make an assumption about the distribution of the response. A further disadvantage of quantile regression is that the response variable must be continuous. This is especially problematic in the case of discrete or binary data, continuous distributions with a probability greater than zero for certain values or when the dependent variable is a proportion. This is different to the GAMLSS approach that also includes those cases. Note that we appraise GAMLSS as a generic framework here, even though it does not yield additional benefits if the distribution has only one parameter such as the binomial or Poisson distribution. Another problem in quantile regression is the issue of crossing quantiles (Bassett and Koenker, 1982). Theoretically, quantiles should be monotonically ordered according to their level such that $\beta_{0,\tau_1} + \beta_{1,\tau_1} x_i \leq \beta_{0,\tau_2} + \beta_{1,\tau_2} x_i$ for $\tau_1 \leq \tau_2$ and all $x_i$, $i = 1, \ldots, n$. Since the regression models are estimated for each quantile separately, this ordering does not automatically enter the model and crossing quantiles can occur especially when the amount of considered quantiles is large in order to approximate the whole distribution. If one assumes parallel regression lines, crossing quantiles can be avoided. However, in this case the application of quantile regression becomes redundant since for each quantile only the intercept parameter shifts while the effect of the explanatory variables would be independent from the quantile level. Therefore, the models rely on the less restrictive assumption that quantiles should not cross for the observed values of the explanatory variables. Strategies to avoid quantile crossing include simultaneous estimation, for example, based on a location scale shift model (He, 1997), on spline based non-crossing constraints (Bondell et al., 2010), or on quantiles sheets (Schnabel and Eilers, 2013). Chernozhukov et al. (2010) and Dette and Volgushev (2008) propose estimating the conditional distribution function first and inverting it to obtain quantiles. However, all of these alternatives require additional steps and most of them cannot easily incorporate an additive structure for the predictors (Kneib, 2013). In empirical research, conventional quantile regression is predominantly used by far. In any case, quantile regression estimates the relationship for certain quantiles separately but does not have a model to estimate the complete distribution. This can be also problematic if measures other than the quantiles such as the standard deviation or Gini coefficient should be analyzed.

In contrast, GAMLSS are consistent models from which any feature of a distribution can be derived. If the assumed distribution is appropriate, GAMLSS can provide more precise estimators than quantile regression especially for the tails of the empirical distribution where data points are scarce. Since we use maximum likelihood for estimation, a variety of related methods and inference techniques that rely on the distributional assumption can be used such as likelihood ratio tests and confidence intervals. As simulation studies in Klein, Kneib and Lang (2015) show bad performance for likelihood-based confidence intervals in certain situations, we will, however, rely on bootstrap inference for the application in Section 3.4. The main drawback of GAMLSS is a potential misspecification but Section 3.4 presents associated model diagnostics to minimize this risk. Besides the methodological differences, quantile regression and GAMLSS expose their benefits in different contexts. Following Kneib (2013), we suggest using quantile regression if the interest is on a certain quantile of the distribution of the dependent variable. On the other hand, the GAMLSS framework is more appropriate if one is interested in the changes of the entire conditional distribution, its parameters and certain distributional measures relying on these parameters, such as the Gini coefficient.

## 3.3 Potentials and pitfalls of GAMLSS for analyzing treatment effects beyond the mean

GAMLSS can be applied to evaluation questions when the outcome of interest is not the difference in the expected mean of treatment and comparison group but the whole conditional distribution and derived distributional measures. Compared to an analysis where the distributional measures are themselves the dependent variable, the great advantage of GAMLSS is that they yield *one* model from which several measures of interest can be coherently derived. In case of income, for example, these measures might be expected income, quantiles, Gini, the risk of being poor etc. Thereby, consistent results are obtained since all measures are based on the same model using the same data. Furthermore, aggregated distributional measures as dependent variables mask the underlying individual information. On the contrary, GAMLSS allows the researcher to estimate (treatment) effects on aggregate measures on the individual level.

When evaluating a program, GAMLSS should be used if the final analysis still includes covariates. In a setting without any covariates, the distribution of the outcome can just be estimated separately (e.g. by plotting the kernel densities) and contrasted. Likewise, quantities derived from these distributions (e.g., the Gini coefficient) could be directly compared between treatment and comparison group. GAMLSS are not required in this case as the central idea of relating all distributional parameters to covariates would become redundant.

After estimating the effects on each distributional parameter, these estimates can be used to calculate the effects on policy-relevant measures or to graphically compare the conditional distributions of the treated and untreated groups. The graphical comparison visualizes where and how the conditional distribution changes due to the treatment.

The GAMLSS framework comprises a wide range of potential distributions and is not bound to the exponential family only such as generalized linear models (GLM). Basically, the dependent variable can take on very different types of distributions as mentioned in Section 3.2.1. For applied researchers or practitioners in impact evaluation, we consider the easy incorporation of mixed distributions as particularly fruitful. When evaluating the effect of a treatment, researchers are often confronted with nonnegative outcomes that have a spike at zero. Regarding count data, an example would be the number of hospital visits with a lot of individuals not having any visit at all. In the case of continuous data, income is a good example

as individuals that do not work have an income of zero. It is common in the evaluation literature and in empirical economics to log transform the income variable in order to meet the normality assumption facilitating easy inference in ordinary least squares (OLS). However, there is an ongoing debate on how to treat values of zero, that is, whether observations can be dropped, replaced by a small positive number, or should not be log transformed at all. While these options might be (arguably) acceptable when there are only few zero valued outcomes, researchers run into problems if this amount is not negligible. As an alternative to commonly applied models to tackle these problems (e.g., the tobit model), zero-adjusted distributions such as the zero-adjusted gamma can be used. This is basically a mixed distribution, with a parameter for the probability of observing a zero and two parameters for the positive, continuous part. Similarly, zero inflated Poisson distributions are a popular choice when modeling count data with a lot of observations at zero. This distribution has two parameters: one for modeling the probability of zero and one for the discrete part.

Another useful distribution that is included in the GAMLSS framework is a distribution for shares. A good example would be if the evaluator wants to analyze if farmers change the composition of land use activities on their fields due to an agricultural intervention. Since shares sum up to one, it is disadvantageous to analyze them in separate regression specifications. For these cases, the Dirichlet distribution provides a suitable distribution. The above examples can be of course analyzed with alternative approaches, we however emphasize the flexibility of GAMLSS in providing a toolbox that can be applied to a wide range of different research problems. The distributions mentioned can be easily employed within the GAMLSS framework and all of them except for the Dirichlet distribution are already implemented in `gamlss` along with other nonstandard distributions. The Dirichlet distribution in a distributional regression framework is currently only available in `BayesX`; see Klein, Kneib, Klasen and Lang (2015) for an application.

Finally, as shown in Section 3.2.2, GAMLSS structure these models in a modular fashion such that several type of effects other than linear ones can be incorporated. This is particularly useful if the relationship between an independent variable and response is nonlinear and better accounted for by splines, if spatial heterogeneities are present, or if panel or hierarchical data are analyzed.

Despite these potentials, it is important to address some limitations regarding model selection and a priori model specification. As the researcher has to select explanatory variables for more than one parameter and a suitable response distribution, uncertainty in estimation can increase yielding invalid $p$-values and possibilities for $p$-hacking open up. Note, however, that there is a trade-off between misspecification by simplifying the model via assuming constant distributional parameters and misspecifying a more complex model. Additionally, a linear regression model is certainly less complex to specify but more limited in its informative value. To reduce the chance for misclassification of more complex GAMLSS, we suggest scrutinizing the model using the criteria and tools for model diagnosis presented in Section 3.4. It is also common in practice to report more than one model to check robustness to model specification.

The second point of a priori model specification is not so much of an issue for most studies relying on observational data when pre-registration is pointless because the data are already available prior to the pre-analysis plan. It is rather related to planned experiments with associated data collection. The superior procedure for experiments is conducting a pre-analysis plan including a hypothesis to be tested, covariates to be included, and an assumption for the response distribution. Specifying covariates for distributional parameters beyond the mean is more difficult than in linear regression; still the same recommendations apply: They can be pre-specified either on theoretical grounds or by using information from previous studies. To some extent, this is also possible for the response distribution. The type of response (continuous, nonnegative, binary, discrete etc.) already restricts the set of possible distributions to choose from. Previous studies might also give hints about the distribution of the response.

To present some examples of beyond-the-mean-measures, we focus in the following on inequality and vulnerability to poverty but a lot more measures can be analyzed using GAMLSS. For example, as Meager (2016) points out, risk profiles of business profits which are important for the functioning of the credit market are based on characteristics of the entire distribution and not only the mean.

### Example: GAMLSS and vulnerability as expected poverty

Ex ante poverty measures such as vulnerability to poverty are an interesting outcome if one is not only interested in the current (static) state of poverty but also in the probability of being poor. Although there are different concepts of vulnerability, see Celidoni (2013) for an overview and empirical comparison of different vulnerability measures, we focus on the notion of vulnerability as expected poverty (Chaudhuri et al., 2002). In this sense, vulnerability is the probability of having a consumption (or income) level below a certain threshold. To calculate this probability, separate regressions for mean and variance of log consumption are traditionally estimated using the feasible generalized least squares estimator (FGLS, Amemiya, 1977), yielding an estimate for the expected mean and variance for each household. Concretely, the procedure involves a consumption model of the form

$$\ln y_i = \beta_0^\mu + \mathbf{x}_i' \boldsymbol{\beta}_1^\mu + \varepsilon_i,$$

where $y_i$ is consumption or income, $\beta_0$ an intercept, $\mathbf{x}_i$ is a vector of household characteristics, $\boldsymbol{\beta}_1$ is a vector of coefficients of the same length and $\varepsilon_i$ is a normally distributed error term with variance

$$\sigma_{e,i}^2 = \beta_0^\sigma + \mathbf{x}_i' \boldsymbol{\beta}_1^\sigma.$$

To estimate the intercepts $\beta_0^\mu$ and $\beta_0^\sigma$ and the vectors of coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_1^\sigma$ the 3-step FGLS procedure involves several OLS estimation and weighting steps. Assuming normally distributed log incomes $\ln y_i$, the estimated coefficients are plugged into the standard normal cumulative distribution function

$$\widehat{\Pr}(\ln y_i < \ln z | \mathbf{x}_i') = \Phi\left(\frac{\ln z - (\hat{\beta}_0^\mu + \mathbf{x}_i' \hat{\boldsymbol{\beta}}_1^\mu)}{\sqrt{\hat{\beta}_0^\sigma + \mathbf{x}_i' \hat{\boldsymbol{\beta}}_1^\sigma}}\right),$$

where $\hat{\beta}_0^\mu + \mathbf{x}_i' \hat{\boldsymbol{\beta}}_1^\mu$ is the estimated mean, $\sqrt{\hat{\beta}_0^\sigma + \mathbf{x}_i' \hat{\boldsymbol{\beta}}_1^\sigma}$ the estimated standard deviation, and $z$ the poverty threshold. A household is typically classified as vulnerable if the probability is equal or larger than 0.5. In contrast to the 3-step FGLS procedure, GAMLSS allow us to estimate the effects on mean and variance simultaneously avoiding the multiple steps procedure. While the efficiency gain of a simultaneous estimation is not necessarily large, its main advantage is the quantification of uncertainty as it can be assessed in one model. In a stepwise procedure, each estimation step is associated with a level of uncertainty that has to be accounted for in the following step. Additionally, GAMLSS provide the flexibility to relax the normality assumption of log consumption or log income.

### Example: GAMLSS for inequality assessment

Although inequality is normally not a targeted outcome of a welfare program, it is considered as an unintended effect since a change in inequality is likely to have welfare implications. To assess inequality, our application in Section 3.4 concentrates on the Gini coefficient but other inequality measures are also applied. In general, we focus on the conditional distribution of consumption or income, that is, the

treatment effects will be derived for a certain covariate combination. In other words, in order to analyze inequality, we do not measure unconditional inequality of consumption or income, for instance, for the entire treatment and comparison group, but inequality given that other factors that explain differences in consumption are fixed at certain values. Thus, for each combination of explanatory variables an estimated inequality measure is obtained which represents inequality unexplained by these variables. The economic reasoning is that differences in consumption or income are not *per se* welfare reducing inequality since those differences might stem from different characteristics or abilities such as years of education. We, however, assess the differences in consumption or income for those with equal or similar education as it is the conditional inequality that is perceived as unfair.

## 3.4 Applying GAMLSS to experimental data

### 3.4.1 General procedure

To demonstrate how the analysis of treatment effects can benefit from GAMLSS, we replicate and extend an evaluation study of a popular intervention and show how a distributional analysis could be implemented step by step.

In particular, we propose the following procedure to implement GAMLSS:

(a) Choose potentially suitable conditional distributions for the outcome variable.

(b) Make a (pre-)selection of covariates according to your hypothesis, theoretical considerations, etc.

(c) Estimate your models and assess their fit, decide whether to include nonlinear, spatial, and/or random effects.

(d) Optionally: Refine your variable selection according to statistical criteria.

(e) Interpret the effects on the distributional parameters (if such an interpretation is available for the chosen distribution), derive the effects on the complete distribution and identify the treatment effect on related distributional measures.

In the following, we apply all of these steps to the Progresa data as used in Angelucci and De Giorgi (2009) to provide a hands on guide on how to use GAMLSS in impact evaluation. The conditional cash transfer (CCT) program Progresa (first renamed Oportunidades and then Prospera) in Mexico is a classical development program. In general, conditional cash transfer programs transfer money to households if they comply with certain requirements. In the case of Progresa, these conditions comprise, for example, children's regular school attendance. CCTs have been popular development instruments over the last two decades and most researchers working in the area of development economics are well familiar with their background and related literature. They thus provide an ideal example for our purpose.

### 3.4.2 Application: Progresa's treatment effect on the distribution

In their study "Indirect Effects of an Aid Program", Angelucci and De Giorgi (2009) investigate how CCTs to targeted, eligible (poor) households affect, among other outcomes, the mean food consumption of both eligible and ineligible (non-poor) households. An RCT was conducted at the village-level and information is available for four groups: eligible and ineligible households in treatment and control villages. Aside from the expected positive effect of the cash transfer on the mean eligible households' food consumption,

Angelucci and De Giorgi (2009) also find a considerable increase of the mean ineligible households' food consumption in the treatment villages. They link the increase to reduced savings among the non-poor, higher loans, and monetary and in-kind transfers from family and friends. The strong economic interrelationships between households within a village presumably result from existing informal credit and insurance markets in the study region. Accordingly, the average program effect on food consumption for the treated villages is larger than commonly assumed when only looking at the poor. Estimating the same relationship using GAMLSS provides important information for the policymakers on the effects within a group, for example, whether conditional food consumption inequality decreases for an average household among the poor (or the non-poor or all households). We will assess the effect on conditional inequality via the Gini coefficient, which is in general defined by

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|y_i - y_j|}{2n\sum_{h=1}^{n} y_i}, \; 0 \leq G \leq 1, \tag{3.3}$$

for a group of $n$ households, where $y_i$ denotes the nonnegative consumption of household $i$. For a given continuous consumption distribution function $p(y)$, which we will estimate via GAMLSS, the Gini coefficient can be written as

$$G = \frac{1}{2\mu}\int_0^\infty \int_0^\infty p(y)p(z)\,|y - z|\,dy\,dz, \tag{3.4}$$

with $\mu$ denoting the mean of the distribution.

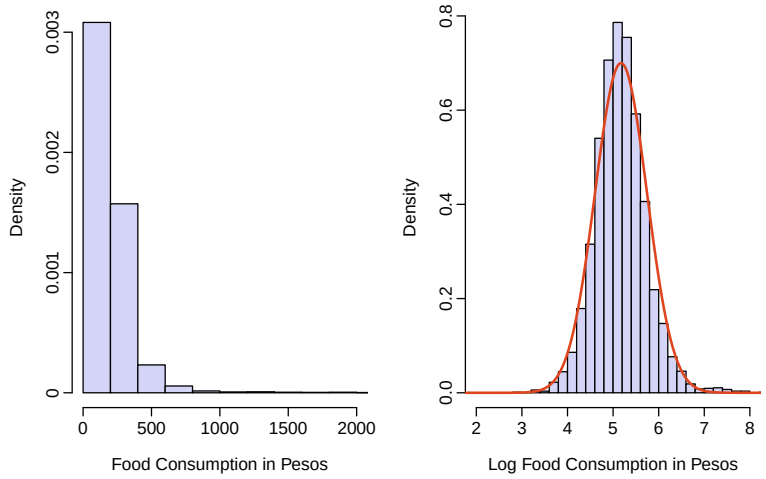
Thus, a positive treatment effect on consumption in one group results in a lower Gini coefficient if all group members benefit equally, as the deviations in the numerator in (3.3) and remain the same, but the denominator increases. An equivalent logic applies to (3.4). However, there might be as well reasons why in one group, for instance among the poor, only the better off benefit and the poorest do not, resulting in higher inequality.

Using GAMLSS, we investigate the program's impact on conditional food consumption inequality measured by the Gini coefficient within the non-poor and poor by comparing the treatment and control groups. In particular, we model food consumption by an appropriate distribution and link its parameters to the treatment variable and other covariates. We obtain estimates for the conditional food consumption distribution for treated and untreated households and the corresponding Gini coefficients. The pairs cluster bootstrap is applied for obtaining an inferential statement on the equality of Gini coefficients; see Section 3.B.2 in the appendix for a description of this bootstrap method.

Furthermore, we investigate the effect of Progresa on global inequality by comparing treatment and control villages, that is, all households in treatment villages are considered as treated and all households in control villages as not treated. Since the average treatment effects found by Angelucci and De Giorgi (2009) are larger for the poor than for the non-poor, a lower food consumption inequality (measured by the Gini coefficient) in the treatment villages is expected. However, a higher Gini could arise if the program benefits are very unequally distributed. Generally, decreasing inequality is an expected, even though often not explicitly mentioned and scrutinized target of poverty alleviation programs and considered to be desirable, especially in highly unequal societies such as Mexico.

In the following, we will therefore investigate the treatment effect on food consumption inequality for three groups: the ineligibles, the eligibles and all households (with those located in a treatment village considered to be treated and vice versa). In particular, we refer to Table 1 in Angelucci and De Giorgi (2009) and restrict our analyses to the most interesting sample collected in November 1999 and the

Figure 3.1: Distribution of food consumption and log food consumption




more powerful specifications including control variables. Generally, we rely on (nearly) the same data and control variables as Angelucci and De Giorgi (2009). Minor amendments for estimation purposes include the removal of households which reported no food consumption and "no answer" categories from categorical variables. The resulting sample size reduction amounts to less than 1% in all samples. In comparison to Angelucci and De Giorgi (2009), we obtained very similar point estimates and significance statements even with our slightly amended sample. Following them, we also remove observations with a food consumption level of more than 10,000 pesos per adult equivalent. Along the steps described in Section 3.4.1 we will show in detail how to apply our modeling framework to the group of ineligibles which are also the main focus group of Angelucci and De Giorgi (2009). Result tables on the remaining two groups are reported and interpreted, whereas a description of the exact proceeding is dropped for the sake of brevity. All necessary software commands and the dataset are available online. The corresponding software code can be downloaded from `https://www.uni-goettingen.de/de/511092.html`, whereas the dataset is available on `https://www.aeaweb.org/articles?id=10.1257/aer.99.1.486`.

## Choice of potential outcome distributions

The distribution of the outcome variable often gives some indication about which conditional distributions are appropriate candidates. However, the (randomized) normalized quantile residuals (Dunn and Smyth, 1996) are the crucial tool to check the adequacy of the model fit and thus the appropriateness of the chosen distribution, as discussed below.

The histogram of the dependent variable in the left panel of Figure 3.1 shows a heavily right-skewed distribution.

The logarithm of the dependent variable in the right panel of Figure 3.1 somewhat resembles a normal distribution such that the log-normal distribution appears to be a reasonable starting point. It has the additional advantage that it also renders easily interpretable effects of the explanatory variables on the mean and variance of the dependent variable, at least on the logarithmic scale. As a more flexible alternative, we will also consider the three-parameter Singh-Maddala that is also known as Burr Type XII distribution and capable of modeling right-skewed distributions with fat tails, see Kleiber and Kotz (2003) for details. Note that the three parameters of the Singh-Maddala distribution do not allow a direct interpretation of effects on moments of the distribution.

## Preliminary choice of potentially relevant covariates

We select the same covariates as in Angelucci and De Giorgi (2009) and relate all of them to all parameters of our chosen distribution. In particular, the model contains nine explanatory variables per parameter: Aside from the treatment variable, these are six variables on the household level, namely poverty index, land size, the household head's gender, age, whether she/he speaks an indigenous language and is illiterate, as well as a poverty index and the land size as variables on the locality level. For the model relying on a log-normal distribution, two parameters $\mu$ and $\sigma$ are related to these variables,

$$\log(\mu_i) = \beta_0^\mu + T_i\beta_T^\mu + \mathbf{x}_i'\boldsymbol{\beta}_1^\mu, \tag{3.5}$$

$$\log(\sigma_i) = \beta_0^\sigma + T_i\beta_T^\sigma + \mathbf{x}_i'\boldsymbol{\beta}_1^\sigma, \tag{3.6}$$

where $T_i$ is the treatment dummy, $\beta_T^\mu$ and $\beta_T^\sigma$ are the treatment effects on the parameters $\mu$ and $\sigma$, respectively, $\mathbf{x}_i$ is a vector containing the values of the remaining covariates for household $i$ and $\boldsymbol{\beta}_1^\mu$ and $\boldsymbol{\beta}_1^\sigma$ are the corresponding coefficient vectors of the same length. In the specification relying on the three-parameter Singh-Maddala distribution, where $\mu$ and $\sigma$ are modeled as in (3.5) and (3.6), respectively, an additional parameter $\tau$ is linked to the nine explanatory variables,

$$\log(\tau_i) = \beta_0^\tau + T_i\beta_T^\tau + \mathbf{x}_i'\boldsymbol{\beta}_1^\tau,$$

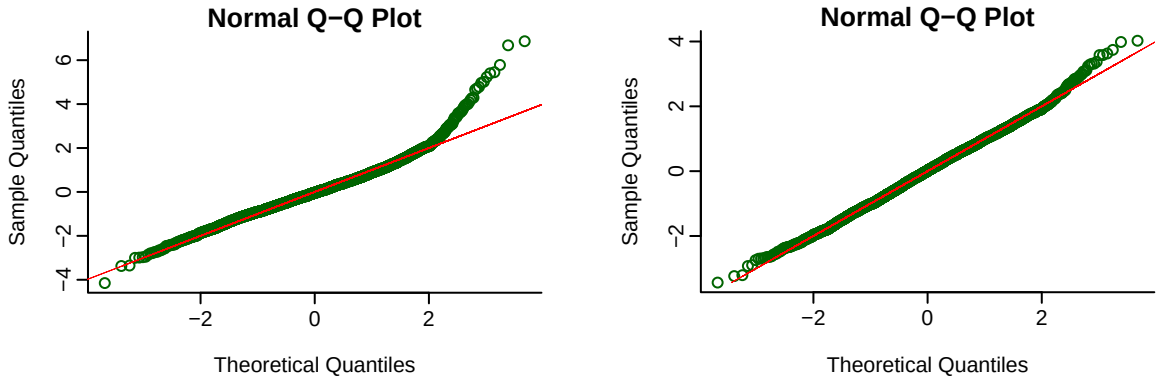
resulting in the considerable amount of 30 quantities to estimate as each parameter equation includes an intercept. This is, however, still a moderate number considering the sample size of more than 4,000 households in the sample of ineligibles and even less problematic for the sample of eligibles with about 10,500 observations and the combined sample. In general, if the sample size is large, it is advisable to relate all parameters of a distribution to all variables which potentially have an effect on the dependent variable and its distribution, respectively. Exceptions may include certain distributions such as the normal distribution when there are convincing theoretical arguments why a variable might affect one parameter such as the mean but not another one such as, for example, the variance. For smaller sample sizes, higher order parameters such as skewness or kurtosis parameters may be modeled in simpler fashion with few explanatory variables.

## Model building and diagnostics

The proposed models are estimated using the `R` package `gamlss`, see Stasinopoulos and Rigby (2007), Stasinopoulos et al. (2017) and the software code attached to this paper for details. The adequacy of fit is assessed by some statistics of the normalized quantile residuals, introduced by Dunn and Smyth (1996). As a generic tool applicable to a wider range of response distributions than deviance or Pearson residuals, these residuals were shown to follow a standard normal distribution under the true model. In Figure 3.2a and Table 3.1 it can be seen that both q-q plot and statistics reveal that the log-normal distribution might be an inadequate choice for modeling the consumption distribution as especially the overly large coefficient of kurtosis, which should be close to 3, and the apparent skewness of the normalized quantile residuals, visible in the plot, suggest a distribution with a heavier right tail.

In contrast, a model relying on the Singh-Maddala distribution yields a much more satisfying diagnostic fit (see Figure 3.2b and Table 3.1). The q-q plot does not show severe deviations from the standard normal distribution, which is confirmed by the summary measures of the quantile residuals. More specifically,

Figure 3.2: Diagnosis plots for the model based on (a) the log-normal distribution and (b) the Singh-Maddala distribution




(a) (b)

Table 3.1: Summary of the quantile residuals for the model based on the log-normal distribution and Singh-Maddala distribution

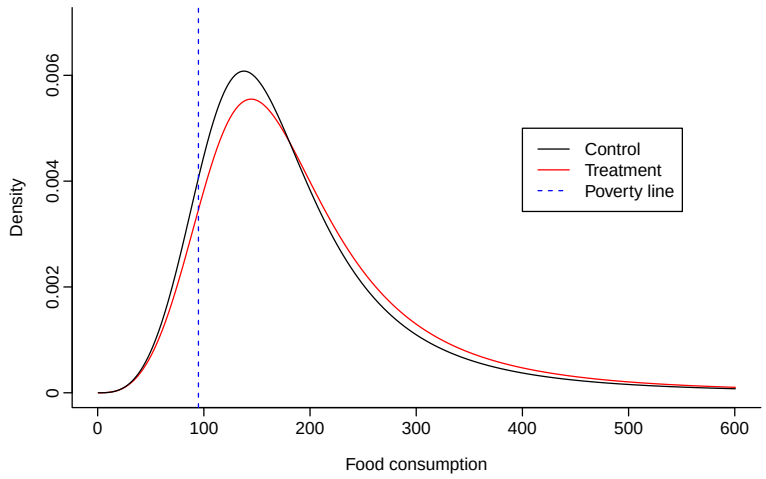
| | Log-normal | Singh-Maddala |
|---|---|---|
| Mean | -0.000091 | -0.001102 |
| Variance | 1.000235 | 0.998379 |
| Coef. of Skewness | 0.701639 | 0.060098 |
| Coef. of Kurtosis | 6.016006 | 3.115085 |
| Filliben Correlation Coef. | 0.984499 | 0.999201 |

Notes: A good fit is indicated by values close to 0, 1, 0, 3 and 1 for mean, variance, skewness, kurtosis, and Filliben correlation coefficient, respectively.

the Filliben correlation coefficient (measuring the correlation between theoretical and sample quantiles as displayed in the q-q plot) is almost equal to 1, the coefficient of skewness is now close to 0 and the coefficient of kurtosis close to 3. Additionally, the mean and the variance do not deviate much from their "desired" values 0 and 1, respectively.

Consequently, the Singh-Maddala distribution is an appropriate choice here for modeling consumption. Other diagnostic tools, as described in Stasinopoulos and Rigby (2007), can be applied as well. In any case, well-fitting aggregated diagnostics plots and numbers do not entirely protect against model misspecification and wrong assumptions. Substance knowledge is sometimes required to detect more subtle issues. In their application, Angelucci and De Giorgi (2009) cluster the standard errors at the village level as some intra-village correlation is likely to occur. In a heuristic approach, we regress the quantile residuals of the model above on the village dummies and obtain an adjusted $R^2$ of about 10% and a very low $p$-value for the overall $F$-Test. This suggests unobserved village heterogeneity which we account for by applying a pairs cluster bootstrap procedure to obtain cluster-robust inference. Alternatively, random effects could be applied to model unexplained heterogeneity between villages. We use the same covariates as in Angelucci and De Giorgi (2009). Following them, we refrain from including nonlinear covariate effects in our model specification. As the model diagnostics indicate a reasonable fit and we are not particularly interested in the effects of the continuous covariates, there is no necessity to apply nonparametric specifications here. Nevertheless, we ran a model with nonparametric covariate effects and obtained very similar results. Generally, we advocate the use of nonparametric specifications, for example via penalized splines, for most continuous covariates. Details on when and how to use penalized splines can be found in Fahrmeir et al. (2013) and Wood (2006).

Figure 3.3: Estimated conditional distributions for an average household




## Variable selection

A comparison between different models, for instance between our model of choice from above and more parsimonious models, may be done by the diagnostics tools described in the previous subsection. Alternatively and additionally, statistical criteria for variable selection may be used, see Wood et al. (2016) for a corrected Akaike Information Criterion for GAMLSS. Moreover, boosting is a valuable alternative especially for high-dimensional models (Mayr et al., 2012). An implementation can be found in the `R` package `gamboostLSS` (see Hofner et al., 2016, for a tutorial with examples), yet the set of available distributions is somewhat limited. Here, we retain all variables in the model in order to stay close to the original study.

## Reporting and interpreting the results

GAMLSS using the Singh-Maddala distribution relate three parameters (via link functions) nonlinearly to explanatory variables but do not yield an immediate interpretation of the coefficient estimates on distributional parameters such as the mean. Yet, it is straightforward to compute marginal treatment effects, that is, the effect of the treatment fixing all other variables at some specified values, on the mean and variance as well as on other interesting features of an outcome distribution, such as the Gini coefficient or the vulnerability as expected poverty. The latter we define as the probability of falling below 60% of the median food consumption in our sample (which corresponds to about 95 Pesos). Finally, $t$-tests and confidence intervals can be calculated for testing the presence of marginal treatment effects on various measures.

The results in Table 3.2 show point estimates and 95% bootstrap percentile intervals of marginal treatment effects for an average household, that is, treatment effects evaluated at mean values for the other continuous explanatory variables and modes for categorical variables (for simplicity, we henceforth refer to the term “at means”) on various distributional measures. The expected significant positive treatment effect on the mean of the dependent variable is found and can be interpreted as follows: For an average household, the treatment induces an expected increase in food consumption of about 16.232 pesos per adult equivalent. Although associated with large confidence intervals including zero, the effect on the variances is also positive, indicating a higher variability in the food consumption among the ineligibles in the treatment villages. The Gini coefficient is as well slightly bigger in treatment villages and the

Table 3.2: Treatment effects for ineligibles

| | Estimate | Lower Bound | Upper Bound |
|---|---|---|---|
| MTE on mean | 16.232 | 2.350 | 23.273 |
| MTE on variance | 8463.007 | -2659.279 | 16895.497 |
| MTE on Gini coefficient | 0.014 | -0.009 | 0.036 |
| MTE on Atkinson index (e=1) | 0.012 | -0.008 | 0.033 |
| MTE on Atkinson index (e=2) | 0.018 | -0.010 | 0.050 |
| MTE on Theil index | 0.019 | -0.017 | 0.055 |
| MTE on vulnerability | -0.015 | -0.044 | 0.009 |

Notes: Shown are point estimates for marginal treatment effects at means (MTE) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 4,248$.

Table 3.3: Treatment effects for eligibles

| | Estimate | Lower Bound | Upper Bound |
|---|---|---|---|
| MTE on mean | 28.900 | 17.328 | 35.066 |
| MTE on variance | 4550.073 | 1378.806 | 7942.616 |
| MTE on Gini coefficient | 0.007 | -0.006 | 0.023 |
| MTE on Atkinson index (e=1) | 0.006 | -0.005 | 0.020 |
| MTE on Atkinson index (e=2) | 0.012 | -0.007 | 0.033 |
| MTE on Theil index | 0.007 | -0.010 | 0.028 |
| MTE on vulnerability | -0.077 | -0.122 | -0.062 |

Notes: Shown are point estimates for marginal treatment effects at means (MTE) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 10,492$.

confidence intervals do not reject the null hypothesis of equal food consumption inequality (measured by the Gini coefficients) between treatment and control villages. We also report effects on other inequality measures, namely the Atkinson index with inequality parameters e $= 1, 2$ and the Theil index. The results are qualitatively comparable to the effect on the Gini coefficient. To put it differently: There is no evidence that the treatment decreases inequality for an average household among the ineligibles, even though a positive effect on the average food consumption can be found. Furthermore, vulnerability as expected poverty does not change significantly due to the treatment, yet the point estimate indicates a decrease by -0.015, corresponding to an estimated probability of falling below the poverty line of 0.111 for an average household in the control group and the respective probability of 0.096 for an average household in the treatment group. The findings can be illustrated graphically: Figure 3.3 shows the estimated conditional food consumption distributions for an average household once assigned to the treatment and once assigned to the control group: It can be seen that the distribution for the treated household is shifted to the right which corresponds to a higher mean and a lower probability of falling below the poverty line. Moreover, the peak of the mode is somewhat smaller and the right tail in this right-skewed distribution is slightly fatter, resulting in an increased variance and thus higher inequality.

The preceding analyses were conducted for an average household in the sample of ineligibles. Clearly, marginal effects could be obtained for other covariate combinations to investigate how the (marginal) treatment effect looks like for specific subgroups. Even more heterogeneity can be allowed for by including interactions between the treatment variable and other covariates. In general, we recommend computing marginal effects at interesting and well-understood covariate values rather than average marginal treatment effects which mask the heterogeneity of the single marginal effects and could be affected overly strongly by observations that are not of primary interest. However, aggregating marginal treatment effects over all households in the sample is as straightforward as showing the distribution of all these single marginal effects.

Table 3.4: Treatment effects for all people in treatment villages

| | Estimate | Lower Bound | Upper Bound |
|---|---|---|---|
| MTE on mean | 25.900 | 15.643 | 30.290 |
| MTE on variance | 4828.316 | 804.267 | 7391.555 |
| MTE on Gini coefficient | 0.007 | -0.005 | 0.022 |
| MTE on Atkinson index (e=1) | 0.006 | -0.004 | 0.020 |
| MTE on Atkinson index (e=2) | 0.012 | -0.004 | 0.036 |
| MTE on Theil index | 0.007 | -0.010 | 0.027 |
| MTE on vulnerability | -0.056 | -0.090 | -0.044 |

Notes: Shown are point estimates for marginal treatment effects at means (MTE) and corresponding 95% bootstrap confidence interval bounds based on 499 bootstrap replicates. $n = 14,740$.

Qualitatively the same results emerge for the group of eligibles, as can be seen in Table 3.3. The treatment effects on the mean are even bigger, still the Gini coefficient and other inequality measures do not decline significantly. In contrast, the point estimates rather indicate a slight increase. A significant decrease is observed for the vulnerability as expected poverty.

Of particular interest are the results on the treatment effects on inequality for all households. In Table 3.4, we see no significant decline in food consumption inequality for a household with the average characteristics, a quite sobering result for a poverty alleviation program, even though we find evidence for a smaller vulnerability to poverty due to the treatment. As the graph of estimated conditional distributions looks similar to Figure 3.3, we do not show it here. However, the reasons for the findings are equivalent: The shift of the distribution to the right due to the treatment lowers the risk of falling below the poverty line. Additionally, while unequal benefits from the treatment increase the variability of the consumption, the right tail of the distribution becomes fatter, preventing an arguably desired decline in inequality.

## 3.5 Conclusion

This paper introduces GAMLSS as a modeling framework for analyzing treatment effects beyond the mean. These types of effects are relevant if the evaluator or the researcher is interested in treatment effects on the whole conditional distribution or derived economic measures that take parameters other than the mean into account. The main advantage of GAMLSS is that they relate each parameter of a distribution and not just the mean to explanatory variables via an additive predictor. Hence, moments such as variance, skewness and kurtosis can be modeled and the treatment effects on them analyzed. GAMLSS provide a broad range of potential distributions which allows researchers to apply more appropriate distributions than the (log-)normal. This is especially the case for dependent variables with mass points (e.g., zero savings) or when the dependent variable are shares of a total (e.g., land use decisions). Furthermore, each distribution parameter's additive predictor can easily incorporate different types of effects such as linear, nonlinear, random, or spatial effects.

To practically demonstrate these advantages, we re-estimated the (mean) regression that Angelucci and De Giorgi (2009) applied to evaluate the well-known Progresa program. They found positive treatment effects on poor and non-poor that were larger for the poor (the target group) than for the non-poor. Their findings suggest that the treatment should consequently also decrease inequality within the two groups and within all households. We tested these hypotheses by applying GAMLSS and could not find any evidence for a decline of the conditional Gini coefficient or other inequality measures due to the treatment. An explanation is that the treatment benefited some households distinctly more than others, leading to a higher variance of consumption between households and a higher amount of households

having a considerably high consumption. We thus argue that GAMLSS can help to detect interesting treatment effects beyond the mean.

Besides showing the practical relevance of GAMLSS for treatment effect analysis, this paper bridges the methodological gap between GAMLSS in statistics and popular methods used for impact evaluation in economics. While our practical example considers only the case of an RCT, we also develop frameworks for combining GAMLSS with the most popular evaluation approaches including regression discontinuity designs, differences-in-differences, panel data methods, and instrumental variables in the appendix. We show there further how to conduct (cluster robust) inference using the bootstrap. The bootstrap methods proposed in this paper rely on re-estimation of a GAMLSS model for each bootstrap sample. In cases of large datasets and complex models, such approaches are computationally very expensive. The implementation of a computationally more attractive alternative, maybe in the spirit of the score bootstrap method proposed by Kline and Santos (2012), is desirable.

# Appendix

# 3.A Combining evaluation methods for non-experimental data and GAMLSS

As demonstrated in Section 3.4.1, GAMLSS can be used for the analysis of randomized controlled trials, as those are typically handled within the ordinary regression framework. The same applies to difference-in-differences approaches which only include additional regressors, namely interactions. In the following, we describe how other commonly used evaluation methods and models (see Angrist and Pischke, 2008, for an overview) can be combined with GAMLSS.

## 3.A.1 GAMLSS and panel data models

In the evaluation literature, linear panel data models with fixed or random effects seem to be the preferred choice when individuals are observed over time:

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta}_1 + \alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T_i. \tag{3.7}$$

Here, $i$ denotes the individual and $t$ the time period. The vector of explanatory variables $\mathbf{x}_{it}$ may include a treatment effect of interest, time dummies and control variables. In order to capture unobserved time-invariant factors that affect $y_{it}$, individual-specific effects $\alpha_i$ are incorporated in the model. Commonly, these are modeled as fixed effects if the random effects assumption of independence between the time-invariant effects and the explanatory variables is presumed to fail. The Hausman test is an occasionally used tool to underpin the decision for using fixed effects. Another approach which loosens the independence assumptions was proposed by Mundlak (1978). The idea is to extend the random effects model such that for each explanatory variable which is suspected to be correlated with the random effects, a variable including individual-specific means of that variable is added. If this procedure is done for all explanatory variables, we obtain the model

$$y_{it} = \beta_0 + \mathbf{x}_{it}'\boldsymbol{\beta}_1 + \bar{\mathbf{x}}_i'\boldsymbol{\delta}_1 + \alpha_i + \varepsilon_{it}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T_i, \tag{3.8}$$

where $\alpha_i, i = 1, \ldots, N$, are random effects, $\bar{\mathbf{x}}_i$ is a vector containing the means of the explanatory variables over all $T_i$ time periods for individual $i$, and $\boldsymbol{\delta}_1$ is the vector of associated coefficients. In this specification, the other vector of coefficients $\boldsymbol{\beta}_1$ only includes the effects of the explanatory variables stemming from their variation around the individual-specific means. Hence, $\boldsymbol{\beta}_1$ in (3.8) is equivalent to $\boldsymbol{\beta}_1$ in a fixed effects model according to (3.7).

For nonlinear (additive) panel data models, the same question about the validity of the independence assumption between the random effects and the explanatory variables arises. One can allow for dependence via the Mundlak formulation in the same fashion as described above for linear models, that is, avoiding the explicit inclusion of fixed effects while loosening the independence assumption, see Wooldridge (2002, Ch. 15) for more details. As random effects are an integrated part of the GAMLSS framework, GAMLSS specifications can be easily used to model panel data. Assume that $y_{it}$ follows a distribution that can be described by a parametric density $p(y_{it}|\vartheta_{it1}, \ldots, \vartheta_{itK})$ where $\vartheta_{it1}, \ldots, \vartheta_{itK}$, are $K$ different parameters of the distribution. Then, according to model (3.8), we can specify for each of these parameters an equation

of the form

$$g_k(\vartheta_{itk}) = \beta_0^{\vartheta_k} + \mathbf{x}_{it}'\boldsymbol{\beta}_1^{\vartheta_k} + \bar{\mathbf{x}}_i'\boldsymbol{\delta}_1^{\vartheta_k} + \alpha_i^{\vartheta_k}, \quad i = 1, \ldots, N,\ t = 1, \ldots, T_i,$$

with link function $g_k$, see Sections 3.2.1 and 3.2.2 in the main text for details and extensions.

### 3.A.2 Instrumental variables

Instrumental variable (IV) regression aims at solving the problem of endogeneity bias, for example arising from omitted variables. In this view, an explanatory variable is endogenous, if an unobserved confounder influences the response and is associated with this endogenous variable. That is, we consider the regression specification

$$y = \beta_0 + x_e\beta_e + x_o\beta_o + x_u\beta_u + \varepsilon \qquad \text{with} \quad E(\varepsilon|x_e, x_o, x_u) = 0,$$

where $x_o$ is an observed explanatory variable, $x_e$ the endogenous variable, $x_u$ the unobserved confounder, $\varepsilon$ is an error term and $\beta_o$, $\beta_e$, and $\beta_u$ represent regression coefficients for the observed, endogenous, and unobserved explanatory variable, respectively. However, $x_u$ cannot be observed and thus cannot be included in the model. As $x_u$ is correlated with $x_e$, this violates the assumption that the error term's expectation given all observed variables is zero. As a consequence, the OLS estimator for $\beta_e$ is inconsistent. In order to demonstrate how a suitable instrument can be used to solve this problem in a nonlinear context, we present the approaches developed for generalized linear models (GLM, Terza et al., 2008), and generalized additive models (GAM, Marra and Radice, 2011) and extend them to the GAMLSS context.

#### Instrumental variables in generalized linear models (GLM)

Terza et al. (2008) proposed a two-stage residual inclusion procedure (2SRI) that addresses endogeneity in nonlinear models. In fact, the procedure was already suggested by Heckman (1978) as a means to test for endogeneity. The reason why ordinary two-stage least squares does not work in the nonlinear context is that the expectation of the response variable is associated via a nonlinear function - the link function in GLMs - with the predictor. Due to this function, the unobserved part is not additively separable from the predictor (Marra and Radice, 2011; Amemiya, 1974).

In a GLM framework, we consider the model

$$E(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u) = h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \mathbf{X}_u\boldsymbol{\beta}_u), \tag{3.9}$$

where $\mathbf{y}$ is the outcome variable dependent on $\mathbf{X}_o$, a $n \times S_o$ matrix of observed variables, on $\mathbf{X}_e$, a $n \times S_e$ matrix of endogenous variables, and on $\mathbf{X}_u$ which is a $n \times S_u$ vector of unobserved confounders that are correlated with $\mathbf{X}_e$. Consequently, $\boldsymbol{\beta}_o$ is a $S_o \times 1$ vector, $\boldsymbol{\beta}_e$ a $S_e \times 1$ and $\boldsymbol{\beta}_u$ a $S_u \times 1$ vector of regression coefficients. The function $h(\cdot)$ denotes the response function, or the inverse of the link function.

The model in (3.9) can be written as

$$\mathbf{y} = h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \mathbf{X}_u\boldsymbol{\beta}_u) + \boldsymbol{\varepsilon} \tag{3.10}$$

where the error term $\boldsymbol{\varepsilon}$ is defined as $\boldsymbol{\varepsilon} = \mathbf{y} - h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \mathbf{X}_u\boldsymbol{\beta}_u)$ such that

$$E(\boldsymbol{\varepsilon}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u) = \mathbf{0}.$$

The correlation between $\mathbf{X}_e$ and $\mathbf{X}_u$ is the core of the endogeneity issue at hand. If we were able to observe $\mathbf{X}_u$, consistent estimators for the coefficients in Equation (3.10) could, for example, be obtained via maximum likelihood estimation (under the usual generalized linear model regularity conditions). Without addressing the endogeneity problem, the $\mathbf{X}_u$ would be captured by the error term leading to a correlation between the explanatory variables and the error.

As in the linear case, to tackle this endogeneity problem, we have to find some observed instrumental variables $\mathbf{W}$ that account for the unobserved confounders $\mathbf{X}_u$. The endogenous variables can be related to these instruments and the observed explanatory variables by a set of auxiliary equations

$$\mathbf{x}_{es} = h_s(\mathbf{X}_o\boldsymbol{\alpha}_{os} + \mathbf{W}_s\boldsymbol{\alpha}_{ws}) + \boldsymbol{\xi}_{us}, \qquad s = 1, \ldots, S_e \tag{3.11}$$

where $\mathbf{x}_{es}$ is the $s$-th column vector of $\mathbf{X}_e$, $h_s(\cdot)$ is the response function, $\mathbf{W}_s$ is a $n \times S_{IV_s}$ matrix of IVs available for $\mathbf{x}_{es}$ and $\boldsymbol{\alpha}_{os}$ and $\boldsymbol{\alpha}_{ws}$ are $S_o \times 1$ and $S_{IV_s} \times 1$ vectors, respectively, of unknown coefficients. The number of elements in $\mathbf{W}$ must be equal or greater than the numbers of endogenous regressors and there is at least one instrument in $\mathbf{W}$ for each endogenous regressor. The error term $\boldsymbol{\xi}_{us}$ in this model contains information about the unobserved confounders.

The instrumental variables $\mathbf{W}_s$ in equation (3.11) have to fulfill the following conditions:

(a) being associated with $\mathbf{x}_{es}$ conditional on $\mathbf{X}_o$

(b) being independent of the response variable $\mathbf{y}$ conditional on the other covariates and the unobserved confounders in the true model, that is, $\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u$

(c) being independent of the unobserved confounders $\mathbf{X}_u$.

Terza et al. (2008) propose the following procedure to estimate the models in Equations (3.10) and (3.11):

(a) First stage: Get the estimates $\hat{\boldsymbol{\alpha}}_{os}$ and $\hat{\boldsymbol{\alpha}}_{ws}$ for $s = 1, \ldots, S_e$ from the auxiliary Equation (3.11) via a consistent estimation strategy. One could use maximum likelihood estimation for GLMs here, but nonlinear least squares is also possible. Define

$$\hat{\boldsymbol{\xi}}_{us} = \mathbf{x}_{es} - h(\mathbf{X}_o\hat{\boldsymbol{\alpha}}_{os} + \mathbf{W}_s\hat{\boldsymbol{\alpha}}_{ws}) \qquad \text{for} \quad s = 1, \ldots, S_e.$$

(b) Second stage: Estimate $\hat{\boldsymbol{\beta}}_e$, $\hat{\boldsymbol{\beta}}_o$, $\hat{\boldsymbol{\beta}}_{\hat{\Xi}_u}$ via a GLM or a nonlinear least squares method from

$$E(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \hat{\boldsymbol{\Xi}}_u) = h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \hat{\boldsymbol{\Xi}}_u\boldsymbol{\beta}_{\hat{\Xi}_u}),$$

where $\hat{\boldsymbol{\Xi}}_u$ is a matrix containing $\hat{\boldsymbol{\xi}}_{us}$ from the first stage as column vectors.

The intuition behind this procedure is that $\hat{\boldsymbol{\Xi}}_u$ contains information on the unobserved confounders if the instruments fulfill the above mentioned requirements. Though $\hat{\boldsymbol{\Xi}}_u$ is not an estimate for the effect of the unobserved confounder on the response variable, its contained information can be used to get corrected estimates for the endogenous variable. Since we are eventually interested in $\boldsymbol{\beta}_e$ and not $\boldsymbol{\beta}_u$, we only need the $\hat{\boldsymbol{\Xi}}_u$ as a quantity containing information about $\mathbf{X}_u$ to account for the presence of these unobserved confounders (Marra and Radice, 2011).

### Instrumental variables in generalized additive models (GAM)

Marra and Radice (2011) extend the 2SRI approach to also cover generalized additive models, that allow for nonlinear effects of the explanatory variables on the response variable. A generalized additive model

has the following form

$$\mathbf{y} = h(\boldsymbol{\eta}) + \boldsymbol{\varepsilon}, \qquad E(\boldsymbol{\varepsilon}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u) = 0,$$

where $\mathbf{X}_e = (\mathbf{X}_e^*, \mathbf{X}_e^+)$, $\mathbf{X}_o = (\mathbf{X}_o^*, \mathbf{X}_o^+)$, and $\mathbf{X}_u = (\mathbf{X}_u^*, \mathbf{X}_u^+)$ with matrices containing discrete variables denoted by $^*$ and continuous ones by $^+$. We summarize the discrete parts of the explanatory variables $\mathbf{X}_e$, $\mathbf{X}_o$, and $\mathbf{X}_u$ into $\mathbf{X}^*$ and the continuous parts into $\mathbf{X}^+$, that is, $\mathbf{X}^* = (\mathbf{X}_e^*, \mathbf{X}_o^*, \mathbf{X}_u^*)$ for discrete variables and $\mathbf{X}^+ = (\mathbf{X}_e^+, \mathbf{X}_o^+, \mathbf{X}_u^+)$ for continuous variables. The linear predictor $\boldsymbol{\eta}$ is represented by

$$\boldsymbol{\eta} = \mathbf{X}^*\boldsymbol{\beta}^* + \sum_{l=1}^{L} f_l(\mathbf{x}_l^+), \tag{3.12}$$

where $\boldsymbol{\beta}^*$ is a vector of unknown regression coefficients and $f_l$ are unknown smooth functions of $L$ continuous variables $\mathbf{x}_l^+$. These continuous variables can be modeled, for example, by using penalized splines (Eilers and Marx, 1996). Since we cannot observe $\mathbf{X}_u^*$ and $\mathbf{X}_u^+$, we get inconsistent estimates for all regression coefficients. Provided that suitable instrumental variables can be identified, we can model the endogenous variables with the following set of auxiliary regressions

$$\mathbf{x}_{es} = h_s(\mathbf{Z}_s^*\boldsymbol{\alpha}_s^* + \sum_{j=1}^{J_s} f_j(\mathbf{z}_{js}^+)) + \boldsymbol{\xi}_{us}, \tag{3.13}$$

where $\mathbf{Z}_s^* = (\mathbf{X}_o^*, \mathbf{W}_s^*)$ with corresponding coefficients $\boldsymbol{\alpha}_s^*$ and $\mathbf{Z}_s^+ = (\mathbf{X}_o^+, \mathbf{W}_s^+)$, where $\mathbf{Z}_s^+$ is composed of $\mathbf{z}_{js}^+, j = 1, \ldots, J_s$. Instrumental variables meeting the same requirements mentioned above are again denoted by $\mathbf{W}_s$. The smooth functions $f_j$ for the $J_s$ continuous variables $\mathbf{z}_{js}^+$ include continuous observed variables and continuous instruments. Despite the notation, $f_l$ in (3.12) and $f_j$ (3.13) generally are different functions.

Marra and Radice (2011) propose the following procedure for the 2SRI estimator within the generalized additive models context:

(a) First stage: Get estimates of $\boldsymbol{\alpha}_s^*$ and $f_j$ for $s = 1, \ldots, S_e$ from the auxiliary Equation (3.13) using a GAM method. Define

$$\hat{\boldsymbol{\xi}}_{us} = \mathbf{x}_{es} - h_s(\mathbf{Z}_s^*\hat{\boldsymbol{\alpha}}_s^* + \sum_{j=1}^{J_s} \hat{f}_j(\mathbf{z}_{js}^+)) \qquad \text{for} \quad s = 1, \ldots, S_e. \tag{3.14}$$

(b) Second stage: Estimate

$$E(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \hat{\boldsymbol{\Xi}}_u) = h_s(\mathbf{X}_e^*\boldsymbol{\beta}_e^* + \mathbf{X}_o^*\boldsymbol{\beta}_o^* + \sum_{j=1}^{J} f_j(\mathbf{x}_{jeo}^+) + \sum_{s=1}^{S_e} f_s(\hat{\boldsymbol{\xi}}_{us})), \tag{3.15}$$

where $\mathbf{x}_{jeo}^+, j = 1, \ldots, J$, are column vectors of $\mathbf{X}_{eo}^+ = (\mathbf{X}_e^+, \mathbf{X}_o^+)$.

In this procedure, $f_s(\hat{\boldsymbol{\xi}}_{us})$ accounts for the influence of unmeasured confounders $\mathbf{X}_u$, and we get thus consistent estimates for the observed and the endogenous variables. The set of models in (3.14) and (3.15) can be fitted by using one of the GAM packages in `R`, for example. In simulation studies, Marra and Radice (2011) show good performance of the estimates if the instruments are strong.

## Instrumental variables and GAMLSS

The IV estimation procedure for generalized linear models and generalized additive models can now be transferred to the GAMLSS context. In these models, the response $\mathbf{y}$ follows a parametric distribution with $K$ distributional parameters $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)'$ and density

$$p(\mathbf{y}|\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u) = p(\mathbf{y}|\boldsymbol{\vartheta}(\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u))$$

For each of the parameters, a regression specification

$$\vartheta_k = h_k(\eta^{\vartheta_k})$$

is assumed, where $\eta^{\vartheta_k}$ is the regression predictor. For each of the predictors $\boldsymbol{\eta}^{\vartheta_k}$ considered over all $n$ observations, we assume a semiparametric, additive structure

$$\boldsymbol{\eta}^{\vartheta_k}(\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u) = \mathbf{X}^*\boldsymbol{\beta}^{*,\vartheta_k} + \sum_{l=1}^{L} f_l^{\vartheta_k}(\mathbf{x}_l^+) \tag{3.16}$$

Using the same notation as above, the only difference between the Equations (3.12) and (3.16) is that the predictors are now specific for each of the $K$ parameters of the response distribution. Note that the predictors do not have to include the same variables, though the indexes are dropped here for notational simplicity.

If $\mathbf{X}_e$ and $\mathbf{X}_u$ are correlated, then $\mathbf{X}_e$ is endogenous and estimating (3.16) without considering $\mathbf{X}_u$ leads to inconsistent estimates due to omitted variable bias.

We propose a similar procedure for GAMLSS as the one Marra and Radice (2011) developed for GAMs:

(a) First stage: Same as for the GAM procedure.

(b) Second stage: Instead of a GAM, estimate a GAMLSS with density $p(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \hat{\boldsymbol{\Xi}}_u)$ and predictors

$$\boldsymbol{\eta}^{\vartheta_k} = \mathbf{X}_e^*\boldsymbol{\beta}_e^{*,\vartheta_k} + \mathbf{X}_o^*\boldsymbol{\beta}_o^{*,\vartheta_k} + \sum_{j=1}^{J} f_j^{\vartheta_k}(\mathbf{x}_{jeo}^+) + \sum_{s=1}^{S_e} f_s^{\vartheta_k}(\hat{\boldsymbol{\xi}}_{us}). \tag{3.17}$$

Wooldridge (2014) has shown that the 2SRI estimator can be used to model $p(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \hat{\boldsymbol{\Xi}}_u)$ in the second step once models for $E(\mathbf{x}_{es}|\mathbf{X}_o, \mathbf{W}_s), s = 1, \ldots, S_e$, are estimated and the $\hat{\boldsymbol{\xi}}_{us}$ are calculated.

To apply Wooldridge's insights to our case, assume we can derive control functions $C_s(\mathbf{X}_o, \mathbf{x}_{es}, \mathbf{W}_s), s = 1, \ldots, S_e$, such that

$$p(\mathbf{X}_u|\mathbf{X}_o, \mathbf{x}_{es}, \mathbf{W}_s) = p(\mathbf{X}_u|C_s(\mathbf{X}_o, \mathbf{x}_{es}, \mathbf{W}_s)). \tag{3.18}$$

Here, $C_s(\cdot)$ acts as a sufficient statistic to take account of the endogeneity. For example, if

$$\mathbf{x}_{es}|\mathbf{X}_o, \mathbf{W}_s \sim N(\boldsymbol{\eta}^{\vartheta_k}(\mathbf{X}_o, \mathbf{W}_s), \boldsymbol{\sigma}_{es}^2),$$

then

$$\boldsymbol{\xi}_u = \mathbf{x}_{es} - \mathbf{Z}_s^*\boldsymbol{\alpha}_s^* + \sum_{j=1}^{J_s} f_j(\mathbf{z}_{js}^+)$$

is an appropriate control function in the sense that assumption (3.18) holds. In this case, including the first-stage residuals $\hat{\boldsymbol{\xi}}_u$ in the second stage, as described in the IV procedures above, is justified. The

control function approach is also adopted, for instance, in Blundell and Powell (2004) for binary responses and continuous regressors. Instead of using splines in the first stage, they rely on simpler kernel estimators but advocated the use of more sophisticated methods.

Assumption (3.18) does not hold in general if the model for the endogenous variable is nonlinear (first stage). However, as Terza et al. (2008) and Marra and Radice (2011) have shown, 2SRI still works approximately. Wooldridge (2014) recommended including $\hat{\boldsymbol{\xi}}_u$ nonlinearily and/or interactions with $\mathbf{X}_e, \mathbf{X}_o$ in (3.17) to improve the approximation. Furthermore, a simulation study on different 2SRI settings suggested standardizing the variance of the first stage residuals (Geraci et al., 2018).

The procedure's implementation is similar to the previous one. In the first stage, we estimate a GAM model with one of the available software packages and the second stage is estimated using `gamlss`. That is, while in the first stage the expected mean of the endogenous variables conditional on the other explanatory variables and the instruments are modeled, the distributional part comes only into play in the second stage. The reason is that our interest is on the distribution of the response variable and the first stage serves only as an auxiliary model to account for the endogeneity. In similar contexts, when combining two stage IV estimation and expectile regression, Sobotka et al. (2013) show in simulations that it is sufficient to focus on the conditional means in the first stage. They also outline a bootstrap procedure that we modify to our case and is presented in Section 3.B.3.

### 3.A.3 Regression discontinuity design

In the regression discontinuity design (RDD), see, for example, Imbens and Lemieux (2008) and Lee and Lemieux (2010) for introductions, a forcing variable $X_i$ fully (sharp RDD) or partly (fuzzy RDD) determines treatment assignment. We first consider the sharp RDD case and adopt a common notation for the RDD, as used by Imbens and Lemieux (2008), for example. Let the treatment variable be $T_i$ which equals 1 if $X_i$ is bigger than some cutoff value $c$ and 0 if $X_i < c$. Then, one is typically interested in the average treatment effect on the mean at the cutoff value

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathrm{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathrm{E}[Y_i | X_i = x], \tag{3.19}$$

where $Y_i$ is the dependent variable of interest. The two quantities in (3.19) may be generally estimated by fitting separate regression models for all or a range of data on both sides of the cutoff value and calculating their predictions at the cutoff value. More precisely, the conditional mean functions $\mathrm{E}[Y_i | X_i, X_i > c]$ and $\mathrm{E}[Y_i | X_i, X_i < c]$ are linked to a linear model via a continuous link function (e.g., identity or logit link). Note that the full range of generalized linear models is included in this formulation, so $Y_i$ may be binary, for instance. Hereby, the crucial assumption is the continuity in the counterfactual conditional mean functions $\mathrm{E}[Y_i(0) | X_i = x]$ and $\mathrm{E}[Y_i(1) | X_i = x]$, where $Y_i = Y_i(0)$ if $T_i = 0$ and $Y_i = Y_i(1)$ if $T_i = 1$. Provided that the assumption holds, the limiting values in (3.19) can be replaced by the conditional mean functions evaluated at the cutoff and differences in the conditional means can solely be attributed to the treatment. Equally reasonable, one can assume continuity in the density functions $p[Y_i(0) | X_i = x]$ and $p[Y_i(1) | X_i = x]$. In this case, estimators from a wide range of models on many other quantities of the distribution of $Y_i$ (aside from the mean) can be identified in the sharp RDD framework. One example is given in Bor et al. (2014) who model the hazard rate in a survival regression. Frandsen et al. (2012) derive quantile treatment effects within the RDD. Likewise, the toolbox of GAMLSS can be applied in the sharp RDD. More specifically, assume $Y_i$ follows a distribution that can be described by a parametric density $p(Y_i | \vartheta_{i1}, \ldots, \vartheta_{iK})$ where $\vartheta_{i1}, \ldots, \vartheta_{iK}$ are $K$ different parameters of the distribution. Then, in a simple linear model including only the forcing variable, we can specify for each of these parameters an

equation of the form

$$g_k(\vartheta_{ik}) = \beta_0^{\vartheta_k} + X_i\beta_1^{\vartheta_k}, \quad i = 1, \ldots, N,$$

on both sides of the cutoff, where $g_k$ is the link function.

The inclusion of further pre-treatment (baseline) covariates into the regression models of choice on both side of the cutoffs has been deemed uncritical, as they are not supposed to change the identification strategy of the treatment effect of interest, see, for instance, Imbens and Lemieux (2008) and Lee and Lemieux (2010). Rigorous proofs in Calonico et al. (2018) confirm that, under quite weak assumptions, it is indeed justified to adjust for covariates for the frequently used local polynomial estimators in the sharp and fuzzy RDD.

As the interest lies in estimating the treatment effect at the cutoff value, one critical question in the RDD is on which data and in which specification the regressions on both sides of the cutoff should be conducted. Global functions using all data typically need more flexibility and include data far from the cutoff, whereas local estimators rely on a smaller sample size and require the choice of an adequate sample. The apparently most popular approaches in the literature, namely those by Calonico et al. (2014) and Imbens and Kalyanamaran (2012), use local polynomial regression (including the special case of local linear regression) and thus, a restricted sample. The inherent bandwidth choice is done with respect to a minimized MSE of the estimator for the average treatment effect on the mean. Based on this minimization criterion, a cross-validation approach as originally described in Ludwig and Miller (2007) and slightly amended in Imbens and Kalyanamaran (2012), is a valuable alternative. In principle, such a cross-validation based bandwidth selection may be transferable to a local polynomial GAMLSS. However, if relying on local estimates, we do not propose using one single bandwidth but rather check the variability of the estimates for different bandwidths, as, for instance, done in Imbens and Kalyanamaran (2012, Figure 2). Additional caution is advised with regard to the diminished sample size resulting from local approaches, as the potentially quite complex GAMLSS require a moderate sample size. In general, we consider global approaches accounting for possibly nonlinear relationships (e.g., via penalized splines) at least as useful complements to local estimators. In any case, we strongly advocate the visual inspection of a scatterplot displaying the forcing and the dependent variable as well as a careful diagnosis for the estimated models, for example based on quantile residuals in the case of GAMLSS.

The extension to a fuzzy RDD, where the treatment variable $T_i$ is only partially determined by the forcing variable $X_i$, requires some new thinking, namely the idea of compliers. Let us again assume that an individual is supposed to get the treatment if its value of the forcing variable $X_i$ is above a certain cutoff $c$. Then, a complier is an individual that complies with the initial treatment assignment, that is, an individual that would not get the treatment if the cutoff was below $X_i$ but that would get the treatment if the cutoff was higher than $X_i$. Commonly, the interest now lies in the average treatment effect (on the mean) at the cutoff value for compliers

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \ \mathrm{E}[Y|X_i = x] - \lim_{x \uparrow c} \ \mathrm{E}[Y|X_i = x]}{\lim\limits_{x \downarrow c} \Pr(T_i = 1|X_i = x) - \lim\limits_{x \uparrow c} \Pr(T_i = 1|X_i = x)}, \tag{3.20}$$

where the denominator now includes the probabilities of treatment at both sides near the cutoff. The treatment effect in (3.20) is identified under the continuity assumption described above for the sharp RDD and two additional assumptions:

(a) The probability of treatment changes discontinuously at the cutoff value.

(b) Individuals with $X_i$ who would have taken the treatment if $X_i < c$ would also take the treatment if $X_i > c$ and vice versa.

The first assumption ensures that the denominator in (3.20) does not equal zero (in the sharp RDD, the denominator is by design equal to one). The second assumption, often called the monotonicity assumption, implies that the initial treatment assignment does not have an unintended effect. In other words, individuals do not become ineligible for the treatment or discouraged from taking up the treatment exactly by the initial treatment assignment. We refer to Imbens and Lemieux (2008) for a detailed discussion on the average causal effect at the cutoff value for compliers.

As in the sharp RDD, assuming the continuity assumption for the density functions $p[Y_i(0)|X_i = x]$ and $p[Y_i(1)|X_i = x]$ to hold, the numerator in (3.20) may also contain differences in other quantities aside from the conditional means. The probabilities in the denominator in (3.20) can be estimated separately, for example via a logistic regression of the treatment variable on the forcing variable, see also Wooldridge (2002, ch. 21). All remaining considerations from the sharp RDD carry over to the fuzzy case, indicating that GAMLSS can be applied both in the sharp and the fuzzy RDD.

# 3.B Bootstrap inference

In the following, we first describe very generic bootstrapping strategies to obtain inferential statements in the GAMLSS context (Section 3.B.1). Peculiarities of the models discussed in this paper are described in the Sections 3.B.2–3.B.4. Practical recommendations for diagnosing bootstrap estimates are given in 3.B.5.

## 3.B.1 General strategy

To fix ideas, assume without loss of generality that the quantity of interest is denoted by $\theta$ and represents the marginal treatment effect at the means, namely the treatment effect for an average individual on the Gini coefficient. We consider the parametric bootstrap as the natural choice for a parametric model such as a GAMLSS, although a nonparametric bootstrap is possible as well. The parametric bootstrap works as follows:

(a) A GAMLSS is fitted to the dataset at hand including $n$ observations. Therefore, $n$ estimated distributions for the dependent variable are obtained.

(b) A bootstrap sample is generated by drawing randomly one number from each of these estimated distributions.

(c) The GAMLSS from the first step is re-estimated for the current bootstrap sample. For treated and non-treated individuals, the conditional distributions at mean values for other covariates are predicted. For these distributions, the respective Gini coefficients are computed and their difference is calculated. This difference between the coefficients is the estimated marginal treatment effect at means on the Gini coefficient and is denoted by $\hat{\theta}_b$ for the current bootstrap sample.

(d) The two preceding steps are repeated for many times, say $B$ times.

From the resulting $B$ bootstrap estimates $\hat{\theta}_1, \ldots, \hat{\theta}_B$, bootstrap inference can be conducted in different ways. One option is to perform a $t$-test based on the bootstrap variance

$$\hat{V}_{\text{boot}}[\hat{\theta}] = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\hat{\theta}})^2$$

with $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$. To test for significance of the marginal treatment effect on the Gini, the $t$-statistic

$$t = \frac{\hat{\theta}}{\sqrt{\hat{V}_{\text{boot}}[\hat{\theta}]}} \tag{3.21}$$

can be used, where $\hat{\theta}$ may be the estimate for the marginal treatment effect from the original sample or the mean of all bootstrap estimates.

Alternatively, a bootstrap percentile confidence interval can be computed. For instance, the bounds of a possibly asymmetric 95% percentile bootstrap confidence interval are given by the lower 2.5th and the upper 97.5th percentile of the $B$ bootstrap estimates, $\hat{\theta}_1, \ldots, \hat{\theta}_B$. Whereas the idea and implementation of such a confidence interval are straightforward, generally more bootstrap samples and thus, more computational power are required than in the case of using bootstrapped standard errors as outlined above. More elaborate bootstrap confidence intervals exist. Efron (1987), for example, proposed a bias-corrected and accelerated method that we do not discuss here. We refer to Efron and Tibshirani (1994) and Chernick et al. (2011) for more details on parametric and nonparametric bootstrap methods as well as on different techniques to derive bootstrap confidence intervals and $p$-values.

### 3.B.2 Bootstrap inference for grouped and panel data

For random effects panel data models where individuals are observed over time and more generally for all random effects models where individuals are grouped into clusters, one has to sample the random effects from their assumed distribution in each bootstrap step first. The distributions for the dependent variable for each individual can then be estimated and the bootstrap dependent variables are drawn from the resulting distributions, corresponding to the first two steps described in Section 3.B.1.

A different approach to account for grouping structures are cluster-robust standard errors. Cameron and Miller (2015) give a comprehensive overview on cluster-robust inference, also within the bootstrap machinery. As a method also applicable to nonlinear models, they propose a nonparametric pairs cluster bootstrap to obtain cluster-robust inference. Assume again that the aim is a significance statement on the marginal treatment effect at means on the Gini coefficient and that the sample consists of $G$ clusters or groups. Then, repeat the following procedure $B$ times:

(a) Resample $G$ clusters $(\mathbf{y}_1, \mathbf{X}_1), \ldots, (\mathbf{y}_G, \mathbf{X}_G)$ with replacement from the $G$ clusters in the original sample, where $(\mathbf{y}_g, \mathbf{X}_g)$, $g = 1, \ldots, G$, denote the vector of the dependent variable and the matrix of the explanatory variables, respectively, for cluster $g$.

(b) Run the GAMLSS for the bootstrap sample obtained in step (a) and predict the respective conditional distributions at mean values for other covariates for treated and non-treated individuals. For these distributions, the respective Gini coefficients are computed and their difference is calculated. This difference between the coefficients is the estimated marginal treatment effect at the means on the Gini coefficient and is denoted by $\hat{\theta}_b$ for the current bootstrap sample.

In complete analogy to our elaborations for non-clustered data, a bootstrap $t$-test can be conducted with the denominator in (3.21) now based on the cluster-robust variance estimator

$$\hat{V}_{\text{clu;boot}}[\hat{\theta}] = \frac{c}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\hat{\theta}})^2,$$

where $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$ and $c = \frac{G}{G-1}\frac{N-1}{N-K}$ is a finite sample modification with the number of estimated model quantities denoted by $K$.

Alternatively, bootstrap percentile confidence intervals and tests can be constructed from the bootstrap estimates, see the explanations in Section 3.B.1.

### 3.B.3 Bootstrap inference for instrumental variables

Due to the stepwise approach in IV methods, the estimation uncertainty arising from the first stage has to be accounted for in the second stage. In order to draw inference for IV models, we propose the following procedure:

(a) Conduct a parametric bootstrap with $N_b$ replications as described in 3.B.1 for the first stage model in Equation (3.13).

(b) With $\hat{\boldsymbol{\alpha}}_s^{[k]}, k = 1, \ldots, N_b$, denoting all of the first stage estimates including the estimates for the smooth functions $f$, calculate

$$\hat{\mathbf{x}}_{es}^{[k]} = h(\mathbf{Z}_s\hat{\boldsymbol{\alpha}}_s^{[k]})$$

and

$$\hat{\boldsymbol{\xi}}_{us}^{[k]} = \mathbf{x}_{es} - \hat{\mathbf{x}}_{es}^{[k]}.$$

(c) For the distributional model in the second stage, replace $\hat{\boldsymbol{\xi}}_{us}$ with $\hat{\boldsymbol{\xi}}_{us}^{[k]}$ and proceed as in the general parametric bootstrap procedure described in 3.B.1.

As an alternative to the parametric bootstrap in step 1, a nonparametric bootstrap approach can be applied by drawing bootstrap samples from $\mathbf{x}_{es}$ and $\mathbf{Z}_s$ to get estimates $\hat{\boldsymbol{\alpha}}_s^{[k]}$ of the first stage model.

Let the number of replicates in the second stage be $N_d$, yielding a total of $N_b * N_d$ replicates for the estimates of interest in the second stage. This procedure can be computationally costly if $N_b$ or $N_d$ are chosen to be large. See Marra and Radice (2011) for a computationally more efficient procedure that assumes approximately normally distributed estimators in the first and second stage, respectively.

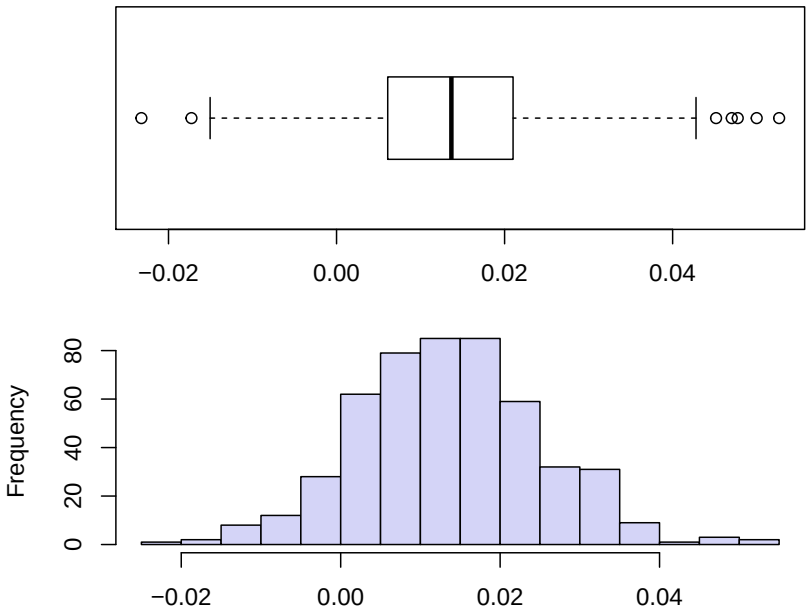
### 3.B.4 Bootstrap inference for RDD

Regressions in the sharp RDD require the estimation of two GAMLSS in each bootstrap sample, namely one on each side of the cutoff value. In the fuzzy RDD, each bootstrap step should also include the re-estimation of the models for the probabilities of the treatment assignment which are chosen to estimate the quantities in the denominator in (3.20). By doing so, the uncertainty of those estimates is included in the resulting standard errors or confidence intervals for the treatment effect of interest.

### 3.B.5 Recommendations for diagnosing bootstrap estimates

Irrespective of the impact evaluation and bootstrap method chosen, but especially in the case of the pairs cluster bootstrap, a thorough inspection of the estimated bootstrap statistics is advisable. If the resulting distribution contains large outliers, one should carefully contemplate disusing or at least amending the currently applied bootstrap procedure. Cameron and Miller (2015) give a more detailed guideline on diagnosing bootstrap estimates. In our example, the distribution of the bootstrap estimates for the

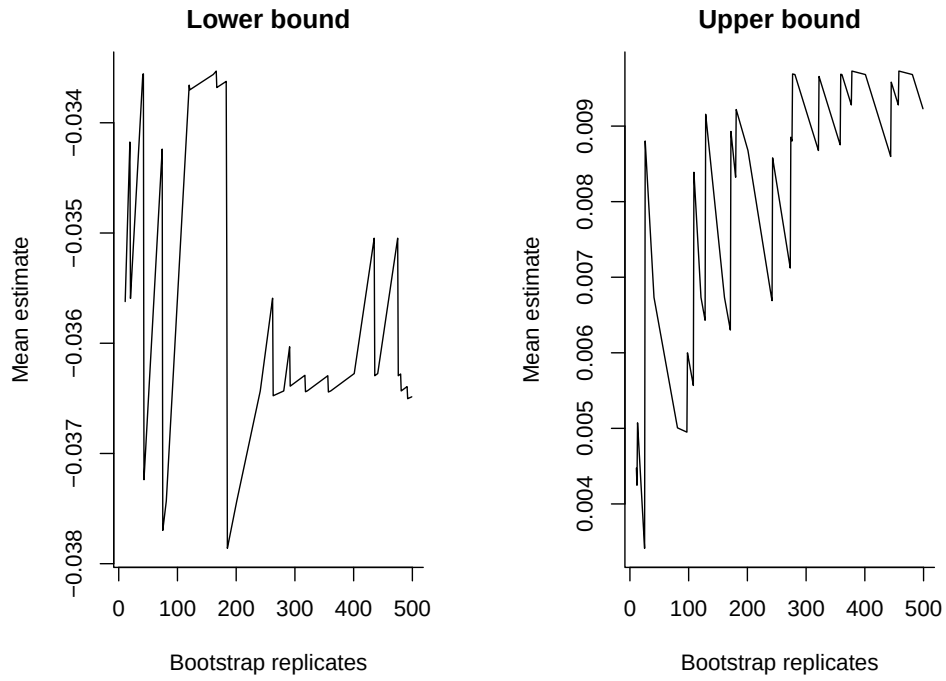
Figure 3.4: Distribution of bootstrap estimates of MTE on Gini




marginal treatment effect at the means on the Gini does not reveal large outliers or severe skewness, as can be seen in the boxplot and the histogram in Figure 3.4.

The question arises of how many bootstrap samples should be generated. Common choices such as $B = 999$ may be applied. Alternatively, inspecting graphically the convergence of the estimated quantities for a growing number of bootstrap samples indicates whether the chosen amount is sufficient. Exemplarily, Figure 3.5 shows the percentile interval bounds for the marginal treatment effects on the Gini in the sample of ineligibles for increasing bootstrap replicates. The chosen bootstrap sample size of $B = 499$ seems to be appropriate as a higher amount of replicates would probably not change the results substantially.

Figure 3.5: Percentile interval bounds for MTE on Gini for increasing bootstrap replicates



# 4 Small area estimation of poverty under structural change

# Small Area Estimation of Poverty under Structural Change

Simon Lange*, Utz Johann Pape†, Peter Pütz‡

**Abstract**

Small area poverty maps allow for the design of policies based on spatial differences in welfare. They are typically estimated based on a consumption survey reporting on poverty and a census providing the spatial disaggregation. This paper presents a new method which allows for the estimation of up-to-date small area poverty maps when only a dated census and a more recent survey are available and predictors and structural parameters are subject to drift over time, a situation commonly encountered in practice. Instead of using survey variables to explain consumption in the survey, the new approach uses variables constructed from the census. The proposed estimator has fewer data requirements and weaker assumptions than common small area poverty map estimators. Applications to simulated data and to poverty estimation in Brazil show an overall good performance.

**Keywords**: Poverty; Population estimates; Censuses; Small Area Estimation

**JEL classification**: D63, I32, R12

*World Bank, Poverty and Equity Global Practice, Africa.

†World Bank, Poverty and Equity Global Practice, Africa. Corresponding author. E-mail: upape@worldbank.org.

‡Economics Department, University of Göttingen.

Authors in alphabetical order. Findings, interpretations and conclusions expressed in this paper are entirely those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments of the countries they represent. The authors would like to thank Pierella Paci and Nobuo Yoshida for valuable comments on earlier drafts.

## 4.1 Introduction

A poverty map is a spatial description of the distribution of poverty in a given country or region. While such a map is useful for policy makers and researchers when small geographic units (e.g., cities, towns, or villages) are discernable, estimates based on household surveys are typically not representative or associated with high uncertainty at such levels of disaggregation. On the other hand, most censuses do not contain information on consumption (or a surrogate such as income or expenditures) required to calculate poverty. To overcome these problems, Elbers et al. (2003), henceforth ELL, developed small area estimation poverty maps, a methodology that can be used to combine information from a detailed household survey with that from a comprehensive census. The general methodology usually consists of two steps, calibration of a statistical model based on survey data and application to the comprehensive census data. In the first step, a multiple linear regression analysis is used to estimate a model of household consumption based on survey data (which includes a consumption module). The explanatory variables in the model are restricted to the subset available in both the survey and the census. These variables are required to be measured in a comparable way both in the census and in the survey.[1] In the second step, the estimated model parameters are applied to census data. The regression model predicts the conditional mean of consumption. Since one is typically also interested in higher moments of the distribution, simulation methods are used to introduce random disturbance. The simulations provide estimates of consumption per capita for every household in the census.

Several criticisms have been raised with regard to the ELL estimator and extensions and alternatives have been discussed. Haslett et al. (2010) propose alternative regression techniques to estimate the survey regression in the first stage. Tarozzi and Deaton (2009) and Molina and Rao (2010) argue that unexplained variation between areas impairs the performance of the ELL estimator as ELL only account for variation between clusters which are nested into areas. While also applying a two-stage approach similar to ELL, Molina and Rao (2010) use area-specific random effects instead of cluster-specific random effects. Moreover, in their empirical Bayes approach they simulate out-of-sample consumption values for the census conditional on the consumption values from the survey. Thus, in contrast to ELL, their simulated census data explicitly includes observed sample information. Das and Chambers (2017) propose another correction for the ELL method which is robust to significant unexplained between-area variability. Their correction relies on the relationship between variance components estimators under the ELL model and a model which additionally contains an area-specific random effect. Marhuenda et al. (2017) discusses the direct application of such a model including cluster-specific and area-specific random effects for poverty mapping via extending the empirical Bayes method of Molina and Rao (2010). Comprehensive discussions on different small area estimation methods can be found in Guadarrama et al. (2016) and Haslett (2016). Still, ELL's is arguably the most frequently used poverty mapping approach combining survey and census data. According to Elbers and van der Weide (2014), it has been applied in more than 60 countries. Some examples for the application of ELL, including fields other than poverty mapping, are Healy et al. (2003), Demombynes and Özler (2005), Elbers et al. (2007), Araujo et al. (2008), Agostini et al. (2010), Bui and Nguyen (2017) and Gibson (2018).

A key assumption for the applicability of ELL is that the distribution of the explanatory variables is the same in both census and survey. This assumption will often be violated if time has passed between data collection for the census and survey, that is, only a dated census and a more recent survey are available, a common situation as censuses are usually conducted less frequently than surveys. Reasons for a violation of this assumption may include demographic trends, migration, natural disasters, and conflicts. If the

[1] Differences in measurement error, coding schemes or even the way the interview was conducted can prevent reasonable harmonization between census and survey variables. See also Tarozzi and Deaton (2009) for a brief discussion.

population parameters, including the regression coefficients, remain unchanged but the distributions of the explanatory variables change over time, ELL results in an outdated poverty map, namely a poverty map at the time of the census. If both the population parameters and the distribution of the explanatory variables change over time, it is not quite clear what is obtained but generally not an up-to-date poverty map.

We propose a different approach that relaxes the discussed assumptions on the explanatory variables. In particular, household characteristics from the census are used to explain consumption values from the survey in the first stage to obtain parameter estimates. In this case, the explanatory variables in the survey are not needed for the estimation and no assumptions on them have to be made. The parameter estimates from the first stage can then be used to predict consumption values using the explanatory variables from the census in the second stage. As it is usually impossible to match households between a census and a survey, the estimation needs to be conducted at a higher geographical level, for instance at the level of census enumeration areas. Throughout this paper, we will refer to the generic term of "clusters" as the lowest level at which census and survey information can be matched. If the assumptions on the explanatory variables hold, this aggregation may worsen the prediction accuracy vis-à-vis ELL, with the magnitude of the loss of precision hinging on the regression model in the first stage. Note that ELL also propose the additional use of census means to explain location effects, namely cluster-specific effects. In this regard, our approach can be considered as a variant of ELL without the use of household-level variables included in both census and survey and without reliance on the associated assumptions. When we refer to the ELL method throughout this paper, we have in mind an estimator that combines survey and census variables at the household-level, the central idea of the approach.

In the case that at least one of the underlying assumptions of ELL is violated, our new approach will still produce up-to-date poverty maps with unbiased poverty estimates. The key assumption we introduce is that *aggregate* household characteristics from the dated census relate to consumption the same way in clusters covered by the recent survey as in clusters not covered by the recent survey. This assumption will hold (on average) if clusters are randomly drawn. Note that a similarly weak assumption has to be made for the applicability of the ELL method if the census and survey are conducted at the same time, namely that household characteristics from the survey relate to consumption the same way in clusters covered by the survey as in clusters not covered by the survey.

In a different scenario, a recent census and only dated survey data may be available. Reliable predictions of poverty measures at the time of the recent census can only be obtained under the additional strong assumption of non-changing structural parameters (including the regression parameters linking explanatory variables to consumption) over time (e.g., Kijima and Lanjouw, 2003). This holds for both ELL and our estimator. If both structural parameters and the distribution of the explanatory variables change over time, ELL results in biased estimates. In contrast, linking census covariate means to predict survey consumption would remain a valid method to generate a poverty map at the time of the survey. In the remainder of this paper, we will focus on the practically more relevant case of a dated census and a recent survey.

Although monitoring poverty over time is of eminent interest to economists (see, for instance, Deaton and Kozel, 2005), little attention has been paid to updating small area estimation approaches which combine dated census and recent survey data. Emwanu et al. (2006) require panel data with one wave collected at the time of the census. While structural changes in the explanatory variables may be detected and tackled by weighting procedures in such a setting, the remaining assumptions of the ELL method as described above are still required. Furthermore, availability of panel data over a longer time span without substantial attrition is rare, especially in developing countries.

The National Statistical Coordination Board of the Philippines (2009) uses explanatory variables deemed time-invariant to estimate intercensal poverty measures. Whether the distribution of variables changes over time is not assessed formally but rather based on *impromptu* assumption. This approach still relies on similar assumptions as the ELL method, even though changes in the distribution of the explanatory variables are ruled out by choosing time-invariant variables. One may also test whether the distribution of potential predictors changed over time and then restrict the set of predictors in the first stage to only those that exhibit no drift.[2] However, severe shocks and extended time periods between survey and census will tend to quickly exhaust the set of viable predictors to do so. And it is exactly in those settings in which the demand for an updated poverty map is likely to be high. Isidro (2010) and Isidro et al. (2016) propose to fit a model on simultaneously collected survey and census data first, for instance by ELL, and update the resulting estimates using a more recent survey. Their Extended Structure Preserving Estimation (ESPREE) approach does not require panel data but contemporaneous surveys and census collection with common variables. The ESPREE method relies on updating multi-way contingency tables which is computationally tractable only for a limited number of categorical explanatory variables and an outcome indicator which is a proportion, for instance the number of people who live below the poverty line. A more general updating procedure is described in Betti et al. (2013). Their propensity score approach also aims at obtaining a covariate distribution in the census as if it was collected at the time of the recent survey. However, the method requires further modeling, including additional assumptions and uncertainty, and a survey collected at the time of the census.

In the remainder of this paper, we show that our proposed method has comparably low data requirements and weak assumptions. Although our outcome variables will be measures of welfare, our method is applicable to a wide range of outcome measures and research questions beyond poverty mapping. Section 4.2 presents the idea of the approach in detail. Section 4.3 describes the properties of the resulting poverty estimator. Simulation studies on artificial and real data are presented in Sections 4.4 and 4.5, respectively. Section 4.6 concludes.

## 4.2 Estimating poverty measures under structural change

Assume that the target population is a village $v$. The quantity of interest is a poverty measure $W$ of the Foster–Greer–Thorbecke (FGT) family (Foster et al., 1984):

$$W_{\alpha v} = \frac{1}{N_v} \sum_{j=1}^{N_v} W_{\alpha v j} \tag{4.1}$$

with

$$W_{\alpha v j} = \left( \tfrac{z - y_{vj}}{z} \right)^{\alpha} I\left(y_{vj} < z\right), \quad \alpha = 0,\ 1,\ 2.$$

Here, $N_v$ is the size of the village population, $y_{vj}$ is the consumption for individual $j$ in village $v$, $z$ is the poverty line and $I(y_{vj} < z)$ is an indicator function which equals one if the consumption of an individual is below the poverty line and zero otherwise. Poverty headcount ratio, poverty gap and poverty severity

[2] This has been suggested for an update of the Bangladeshi poverty maps by researchers from The Bangladesh Bureau of Statistics, The World Bank and The United Nations World Food Programme (2010).

are obtained for $\alpha = 0, 1$ and 2, respectively.[3]

### 4.2.1 The consumption model

In the following, we refer to consumption at the household level since consumption values are usually observed at the level of the household, not the level of the individual. As most household consumption values are unobserved in a village, one needs a model which predicts those values for all households. Let $y_{cht}$ be the consumption of household $h$ in cluster $c$ at time $t$. Then, the model of consideration is

$$\begin{aligned} y_{cht} &= \mathbf{x}'_{c.,t-1}\boldsymbol{\beta} + u_{ch} = \mathbf{x}'_{c.,t-1}\boldsymbol{\beta} + \eta_{ct} + e_{cht}, \quad h = 1, \ldots, H_c, \quad c = 1, \ldots, C, \\ \eta_{ct} &\sim iid\ \mathcal{F}_1\left(0, \sigma_\eta^2\right), \quad e_{cht} \sim iid\ \mathcal{F}_2\left(0, \sigma_e^2\right), \end{aligned} \tag{4.2}$$

which relates the (potentially transformed) consumption variable linearly to a vector $\mathbf{x}_{c.,t-1}$ containing dated census means of covariates over the cluster $c$ from time point $t-1$.[4] The two error components are the cluster effects $\eta_{ct}$ and the household errors $e_{cht}$ which follow the distributions $\mathcal{F}_1$ and $\mathcal{F}_2$, with zero expectation and variances $\sigma_\eta^2$ and $\sigma_e^2$ respectively, and are assumed to be independent of each other. It is possible to allow for heteroscedasticity in the household error by modeling its variance to covariates. Such covariates may include the census means used in the main regression but also higher moments such as the variance. Furthermore, geographic information and the fitted values of the first-stage regression may be used. The ELL method describes one option to model heteroscedasticity within the framework discussed here, while Pinheiro and Bates (2000, ch. 5) provide a more comprehensive discussion.

### 4.2.2 Model estimation based on survey consumption values

In the first stage, model (4.2) is estimated using all household consumption values which are available for the village of interest in the survey. The estimation can be done by weighted or (feasible) generalized least squares.[5] As the estimates are used to predict consumption values for the census, the aim is to find a model with high predictive power. Thus, one should find a model containing only covariates which explain a substantial share of the variation in the dependent variable. Due to averaging over the cluster, means over candidate variables should exhibit variation across clusters.

### 4.2.3 Bootstrapping census consumption data

In the second stage, model (4.2) is used to predict consumption values for each household in the village of interest based on the census. Note that, to be consistent with the first-stage model using the consumption values from the survey, the explanatory variables in the second stage are also averaged within clusters, that is, all households in the same cluster have the same value for each explanatory variable. Using the estimated regression coefficients $\widehat{\boldsymbol{\beta}}$ from model (4.2) yields predictions $\hat{y}_{cht} = \boldsymbol{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}$, that is, predicted

[3] The proposed method is not restricted to measures of the FGT family but applicable to essentially all measures which can be derived from consumption (or any other dependent variable measuring welfare), for instance inequality measures such as the Gini coefficient.

[4] In practice, one could use additional secondary information to explain consumption, for example geographic information which is typically available in poverty mapping exercises. Besides, fixed effects on higher aggregation levels such as counties and time-invariant explanatory variables on the household level could be, in principle, added to the consumption model. As discussed in Section 4.1, we do not assume many time-invariant variables to be available in practice and it is difficult to test if there are any. In this paper, we restrict ourselves to information that is available in the census.

[5] The chosen estimation method depends on whether and how the survey design, potential heteroscedasticity and the clustering nature of the data are taken into account.

conditional means. To account for the deviations of the observed household consumption values from these means, random disturbance terms have to be added by simulation. Assume that the aim is to estimate a poverty measure $W$, where the indices from (4.1) are dropped for notational convenience.

A bootstrap procedure is applied to generate $R$ pseudo censuses and resultant poverty measures:

1. Draw all model coefficients from their respective sampling distribution estimated by the model in the first stage, including regression coefficients, random term variances and possible heteroscedasticity parameters. Multivariate normal distributions with first-stage estimates for the means and the robust variance-covariance matrices accounting for correlation within clusters are used to draw the regression coefficients and potential heteroscedasticity parameters.[6]

2. Conditional on the parameters describing the error components' distributions from the first step, cluster effects and household errors are drawn from their respective distributions. One option is to use a parametric bootstrap, that is, to assume certain parametric distributions for which the estimates from the first stage regression might give some indication. However, a nonparametric bootstrap procedure is a valid alternative or supplement. In this case, a cluster effect can be estimated as the mean of the deviations between observed and predicted values in one cluster, that is, $\widehat{\eta}_{ct} = 1/H_c \sum_h^{H_c} (\widehat{y}_{cht} - \mathbf{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}})$, while the household residuals are computed as those deviations minus the cluster effects, that is, $\widehat{e}_{cht} = (\widehat{y}_{cht} - \mathbf{x}'_{c.,t-1}\widehat{\boldsymbol{\beta}}) - \widehat{\eta}_{ct}$. There are different strategies to draw from these sampling distributions. One may draw with replacement from all estimated cluster effects and all household residuals. Alternatively, the household residuals may be drawn only from the location to which the drawn cluster effect belongs. This strategy generally allows the estimated two error components to be related in a nonlinear way, even though they are by construction (linearly) uncorrelated.

3. Calculate the predicted consumption values for all households and all individuals as well as the poverty measure $\widehat{W}^{(r)}$ derived from those values.

4. Repeat steps 1 to 3 $R$ times.

For the poverty measure $W$, the estimated expected value is then given by

$$\widetilde{\mu} = \frac{1}{R} \sum_{r=1}^{R} \widehat{W}^{(r)}$$

and its estimated variance by

$$\widetilde{V} = \frac{1}{R} \sum_{r=1}^{R} (\widehat{W}^{(r)} - \widetilde{\mu})^2. \tag{4.3}$$

Due to the bootstrap procedure, the variance contains uncertainty from the first-stage model (step 1, referred to as model error in the next section) and the unobservable part of consumption (step 2, referred to as idiosyncratic error in the next section).

[6] One may also assume a distribution for the error components' variances such as the gamma distribution, for example, but in many cases it is reasonable to treat their estimates from the first stage as fixed, especially if the numbers of clusters and households in the survey are large since then there is not much uncertainty in the variance estimators. The household error variance estimator is usually very precise as it is based on the (large) number of households in the survey. The amount of clusters in the survey is smaller but the uncertainty in the variance estimator of the cluster effects is often still negligible. In practice, one may check whether the estimated variances of the error components' variances are small enough in order to treat them as fixed in all bootstrap replications.

## 4.3 Properties of the estimator

In the following, we will investigate the properties of our welfare estimator presented in the previous section.

As described in ELL, the prediction error, the difference between the actual poverty measure $W$ for a target population, say a village, and our estimator $\widetilde{\mu}$ of its expectation $E(W) = \mu$, is given by

$$W - \widetilde{\mu} = (W - \mu) + (\mu - \widehat{\mu}) + (\widehat{\mu} - \widetilde{\mu}) . \tag{4.4}$$

Here, the third component is the computation error which is the difference between our estimator $\widetilde{\mu}$ and its expectation $\widehat{\mu}$. In the following, we assume the computation error to be negligible by applying a sufficiently high number of bootstrap simulations.

The first term on the right-hand side of equation (4.4), $(W - \mu)$, is the idiosyncratic error arising from the unexplained part of consumption of which the poverty measure is a function. Due to the stochastic nature of consumption, the actual poverty measure differs from its expected one. Note that the population in the small area of interest is finite and can be seen as a realization from an infinite population. Hence, all asymptotic results for the idiosyncratic error of the poverty measure from ELL carry over to the new approach presented here: The idiosyncratic error vanishes asymptotically for growing population size, including additional clusters and individuals.

The second part of equation (4.4), $(\mu - \widehat{\mu})$, is the model error, which originates from the estimation of (unknown) population parameters. The expectation of the model error equals zero if the poverty estimator is an unbiased estimator for the expected value of the true poverty measure. Whether this is the case hinges on the regression model selected for the survey data.[7] What is crucial is that the assumptions of zero mean, independence, and homoscedasticity for the error components, namely the cluster effects and the household errors, hold. Likewise, if the error components are assumed to follow certain distributions and these parametric assumptions are used for the generation of simulated census datasets (see Section 4.2.3), they also have to hold. Note that these assumptions may be valid even if dated census data are used for predicting survey consumption values. Thus, one crucial part is the diagnosis of the estimated error components from the first-stage regression. If plots or statistical tests on the estimated cluster effects and household residuals suggest violations of distributional assumptions, one should adjust the model accordingly. More specifically, heteroscedasticity, serial correlation, and non-normality can be detected and accounted for, for instance by choosing different predictor specifications, transforming the dependent variable, or explicit modeling of heteroscedasticity as discussed in Section 4.2.1. The variance of the model error also depends fully on the properties of the first-stage estimators. It typically decreases in survey sample size.

If the assumptions of the ELL method hold and the models are correctly specified, the ELL estimator will usually exhibit a smaller variance of the prediction error than our estimator. The reason is that the latter is a between estimator that ignores variation within clusters. Intuitively, both estimators would only be similarly efficient if the explanatory variables differed distinctly more between clusters than within clusters. In practice, another exception might occur if there are many missing values in the explanatory variables in the survey. Without imputation methods that are subject to estimation uncertainty, the ELL

[7] Note that it is neither intended nor necessary to establish causal or direct effects of explanatory variables on consumption. Thus, the regression coefficients in model (4.2) need not be estimated unbiasedly or consistently with regard to the direct effects of the explanatory variables. In contrast, asymptotic unbiasedness of $\widehat{\mu}$ can be obtained for several models, even if a single parameter in such a model might capture the effect of several correlated variables.

first-stage estimator would be based on a smaller sample than our estimator.

In practice, the variance components of the idiosyncratic and model error are not estimated separately. Rather, the entire variance of the prediction error is obtained from the variation of the simulated poverty estimates in equation (4.3). Hence, under correct distributional assumptions on the random components, the bootstrap procedure allows to draw valid inferences, that is, to build confidence intervals which include the true poverty measure with a predetermined probability. For instance, bootstrap percentile intervals, which can be constructed directly from the bootstrap estimates (see Section 4.2.3), can be used for inference.

Another potential issue in practice is multicollinearity. Note that the fundamental unit of the predictors in the first stage is a cluster, not a household, and that the number of parameters that can be included in (4.2) is hence restricted to the number of clusters. However, household budget surveys that are used to estimate poverty incidence typically cover 500 clusters or more, with some covering substantially more. Hence, we believe that our estimator could be based on a moderate number of regressors that would be sufficient to accurately predict household consumption which is assumed to differ between clusters.[8]

## 4.4 Simulation experiments

A simulation study is conducted to compare the performance of our approach, ELL, and a purely survey-based estimator in predicting FGT poverty measures. We focus on the poverty headcount ratio $W_0$ and the poverty gap $W_1$ with three generic poverty lines that render 25%, 50%, and 75% of the population poor. The simulation setting is based on Tarozzi and Deaton (2009). In particular, the target population in the census is a village with $N = 15,000$ households, divided into 150 clusters $k_c \in \{1, \ldots, 150\}$, each of size 100. In each simulation run, an artificial household survey is drawn from the census by selecting randomly ten households from 100 randomly selected clusters. First, both census and survey are generated by the following process with homoscedastic errors:

$$
\begin{aligned}
&y_{ch} = 25 + x_{ch} + \eta_c + e_{ch} \\
&x_{ch} = 0.01 k_c - t_{ch}, \quad k_c \in \{1, \ldots, 150\}, \quad t_{ch} \sim U(0,1), \\
&\eta_c \sim N(0, 0.01), \quad e_{ch} \sim N(0, \sqrt{2}).
\end{aligned}
$$

Note that the explanatory variable is generated so that it differs in expectation between clusters. Such a situation with large and systematic differences in the averages of covariates across clusters (e.g., average levels of education or dwelling characteristics) is frequently observed in practice. This setting is ideal for the ELL method, which exactly models the data generating process. A linear regression based on the target population yields an $R^2$ of 0.55 while the new method with an $R^2$ of 0.08 has considerably lower explanatory power.

A second setting mimics a real-word situation where the census is dated and a more recent household survey (with an underlying true census which is not observed) is available. Here the model which explains consumption in the same way as the first setting for both the census and the survey, but the explanatory variable for the more recent survey is generated by

[8] One commonly used rule-of-thumb is to restrict the number of predictors to the square root of observations. While our results in Sections 4.4 and 4.5 are based on 100 clusters and less than ten variables, 500 clusters would allow the analyst to base the first-stage estimation on more than 20 census averages (or other summary statistics computed at the cluster-level).

Table 4.1: *Monte Carlo simulation setting 1 - simultaneous census and survey collection, some variation in the explanatory variable between clusters*

| | | New estimator | | | ELL estimator | | | Survey est. |
|---|---|---|---|---|---|---|---|---|
| | True value | Bias | RMSE | Coverage | Bias | RMSE | Coverage | RMSE |
| $W_0(.25)$ | 0.2500 | 0.0025 | 0.0121 | 0.9800 | 0.0017 | 0.0081 | 0.9833 | 0.0137 |
| $W_0(.50)$ | 0.5000 | 0.0062 | 0.0146 | 0.9767 | 0.0058 | 0.0102 | 0.9633 | 0.0159 |
| $W_0(.75)$ | 0.7500 | 0.0028 | 0.0113 | 0.9800 | 0.0036 | 0.0084 | 0.9600 | 0.0144 |
| $W_1(.25)$ | 0.0094 | 0.0000 | 0.0007 | 0.9500 | -0.0000 | 0.0005 | 0.9700 | 0.0007 |
| $W_1(.50)$ | 0.0240 | 0.0002 | 0.0012 | 0.9833 | 0.0001 | 0.0008 | 0.9800 | 0.0012 |
| $W_1(.75)$ | 0.0473 | 0.0003 | 0.0015 | 0.9800 | 0.0002 | 0.0010 | 0.9900 | 0.0015 |

$W_\alpha(r)$ denotes the respective FGT measure for a poverty line that renders a share $r$ of the population poor. The RMSE is the root of the mean squared deviations of the respective estimates from the true value over 300 replications. Coverage rates are calculated for 95% bootstrap percentile intervals.

$$x_{ch} = 0.01k_c, \quad k_c \in \{1, \ldots, 150\},$$

where the sampled 100 clusters in the survey have the same values for $k_c$ as they have in the dated census. For both estimators, the $R^2$ obtained from the first-stage regression for all generated surveys is on average similar to the $R^2$ based on the census in the first setting.

The estimators purely based on the survey simply plug in the observed consumption values from the survey into the FGT measures. Note that in both settings, these estimators have desirable properties as the surveys are representative of the respective village population at the time of data collection. In real-word situations, however, a survey is not necessarily representative at the village-level.

All results are based on 300 Monte Carlo replications with 500 bootstrap census datasets generated in each replication for the two methods which use census data. The bootstrap procedure to sample the error components applies a simple nonparametric version, that is, both cluster effects and household errors are independently sampled with replacement from their sample analogs from the first-stage regression. See Section 4.2.3 for details.

In the first setting, the root mean squared error is, as expected, smallest for the ELL method, followed by our estimator and an estimator solely based on the survey (Table 4.1). Although the $R^2$ from the first-stage regression for the ELL method is seven times as large as for our new method, the root mean squared errors only differ by a factor of about 1.5 or two-thirds, respectively. The coverage rates of the two methods are close to the nominal one of 95% and the bias is negligible.

In the second and more interesting setting, the ELL method naturally is the worst in terms of prediction and generates invalid confidence intervals (Table 4.2). The upward bias originates from the data generating process above: as the expected values of $x_{ch}$ and thus $y_{ch}$ are larger in the recent survey and its underlying population than in the dated census, using the dated census data to predict current poverty statistics necessarily underestimates the current values of $y_{ch}$ and hence overestimates the magnitude of poverty. In contrast, the new method yields valid confidence intervals. It also results in a lower mean squared error in comparison to the purely survey-based estimate since additional census information is exploited. The last result typically holds on average if the model assumptions are fulfilled (as it is the case in this simulation setting) and census and survey size differ distinctly. The latter is often true in practice.[9]

[9]Note that under the stated conditions, our estimator performs better only in predicting the true value on average. In a single sample, the pure survey mean is superior to our approach if the sample mean is by chance equal or very close to the census mean. An extreme example includes the limiting case in which the recent survey is equal to the underlying census. Then, the survey mean is trivially the census mean, that is, there is no error at all. But our new method is still prone to idiosyncratic and (small) simulation error, even under correct model specification.

Table 4.2: *Monte Carlo simulation setting 2 - dated census and recent survey, some variation in the explanatory variable between clusters, explanatory variable changes over time*

| | | New estimator | | | ELL estimator | | | Survey est. |
|---|---|---|---|---|---|---|---|---|
| | True value | Bias | RMSE | Coverage | Bias | RMSE | Coverage | RMSE |
| $W_0(.25)$ | 0.2500 | -0.0035 | 0.0128 | 0.9533 | 0.1186 | 0.1190 | 0.0000 | 0.0155 |
| $W_0(.50)$ | 0.5000 | 0.0040 | 0.0144 | 0.9833 | 0.1374 | 0.1377 | 0.0000 | 0.0166 |
| $W_0(.75)$ | 0.7500 | -0.0011 | 0.0112 | 0.9867 | 0.0925 | 0.0927 | 0.0000 | 0.0154 |
| $W_1(.25)$ | 0.0089 | -0.0001 | 0.0007 | 0.9367 | 0.0065 | 0.0066 | 0.0000 | 0.0008 |
| $W_1(.50)$ | 0.0234 | -0.0002 | 0.0011 | 0.9767 | 0.0115 | 0.0116 | 0.0000 | 0.0012 |
| $W_1(.75)$ | 0.0456 | -0.0001 | 0.0014 | 0.9833 | 0.0154 | 0.0155 | 0.0000 | 0.0016 |

$W_\alpha(r)$ denotes the respective FGT measure for a poverty line that renders a share $r$ of the population poor. The RMSE is the root of the mean squared deviations of the respective estimates from the true value over 300 replications. Coverage rates are calculated for 95% bootstrap percentile intervals.

## 4.5 Application to census data from Brazil

In order to test the proposed method in a real-world example, we use data extracts from the 2000 and 2010 Brazilian censuses provided by the Integrated Public Use Micro Sample (IPUMS, Minnesota Population Center, 2017), the preferred basis of welfare measurement in developing countries. Both censuses include information about monthly income at the level of the individual. In addition, the datasets provide information that is potentially useful in explaining incomes, including the location in which the household resides (urban / rural), the number of household members, ownership of specific assets, and employment status. This allows us to generate artificial surveys from the more recent census and predict income by dated census data. The poverty measures derived from the predicted income values can then be compared to the true ones based on the entire recent census.

The datasets are extracts from the respective censuses. Roughly ten million individuals are included in each dataset, corresponding to 6 and 5 percent of the population in 2000 and 2010, respectively. The country is divided into 25 states and 1,980 municipalities. These municipalities constitute the smallest geographical unit which can be matched between 2000 and 2010. Accordingly, we consider them as clusters in the terminology used in the previous sections. Thus, we use averages over municipalities for the 2000 census to predict household incomes in 2010. Household incomes are calculated as the sum of individual incomes of all household members, adjusted for the household size according to the OECD-modified scale.[10] The poverty line is set to \$5.5 in 2011 PPP per person and day.[11] For the sake of illustration, we focus on one single Brazilian state, Minas Gerais. In comparison to other states, it features a large number of municipalities (282) which we can match over the two censuses. The datasets comprise 303,134 and 359,051 observed households in 2000 and 2010, respectively, with full information on the used variables. Maintaining the ratio of number of households, we sample randomly about 18,188 households (year 2000) and 21,543 (year 2010) from the respective censuses and treat the resulting datasets as new censuses. The reason for that is not only computational convenience but also the fact that the state of Minas Gerais is the small area of interest and should therefore exhibit a population size similar to common empirical applications in small area estimation. The true headcount ratios in these artificial censuses change substantially over time, from 0.27 percent in 2000 to 0.11 percent in 2010.

[10] `http://www.oecd.org/eco/growth/OECD-Note-EquivalenceScales.pdf`.

[11] The World Bank calculates poverty rates at three poverty lines for Brazil, see `http://databank.worldbank.org/data/download/poverty/B2A3A7F5-706A-4522-AF99-5B1800FA3357/9FE8B43A-5EAE-4F36-8838-E9F58200CF49/60C691C8-EAD0-47BE-9C8A-B56D672A29F7/Global_POV_SP_CPB_BRA.pdf`. We chose the highest one since otherwise there are very few households below the other two poverty lines in both years. Our main aim is to illustrate the method's applicability even in settings in which the time span between the datasets is large and relevant changes in the welfare status have occurred over time.

Table 4.3: *Regression results - new estimator using all households from 2010 census*

| Dependent variable: Income | Coefficient estimate | 95% confidence interval |
|---|---|---|
| Phone availability | 0.448 | [0.318; 0.579] |
| Unemployed | -0.518 | [-0.668; -0.367] |
| Urban | 0.233 | [0.126; 0.340] |
| Educational level | 0.335 | [0.248; 0.422] |
| Household members | -0.159 | [-0.188; -0.130] |
| Constant | 2.655 | [2.449; 2.861] |
| Number of census households | 21,543 | |
| Number of municipalities | 282 | |
| $R^2$ | 0.0950 | |

As variables with sufficient variation between municipalities and power to explain variation in income we use the location (share of urban households), the average number of household members, the share of households owning a phone as well as the unemployment rate and the average educational level. The latter is based on the level of schooling completed (measured on a four-point scale) by the person with the highest educational attainment in the household. When all households from the 2010 census are used, a linear regression with these explanatory variables yields an $R^2$ of 0.095. The estimates of the regression coefficients can be found in Table 4.3. We also added squares of the variables, interactions and many other variables to this simple model without obtaining a substantially higher predictive ability measured by the Akaike Information Criterion. The estimated cluster effects variance in a linear mixed effects model based on the 2010 census is 0.02 and small compared to the estimated household residual variance of 0.88.

We draw artificial surveys from the 2010 census by first sampling randomly without replacement 100 municipalities and then sampling without replacement 10 households randomly from each of those municipalities, resulting in an overall survey sample size of 1,000 households. As the number of households differs between municipalities, the estimation at the first stage has to account for these differences by using appropriate weights. Note that this requires knowledge of the number of households in the municipalities at the time of the survey. In practice, when no recent census is available, the number of households at the cluster level can be obtained from a listing exercise which is usually also needed for the sampling scheme for the household survey.

We use a weighted linear regression in the first stage. Means of the explanatory variables over municipalities for the year 2000 are used to explain household per capita income in 2010. To remove apparent right-skewness in the dependent variable, a log-transformation is applied after adding one to the household income values. The latter is done due to the non-negligible amount of zero income values.[12]

In the second-stage bootstrap procedure, the regression coefficients are sampled from a multivariate normal distribution where the expected values and the cluster-robust variance-covariance matrix are the first stage estimates. The error components are generated by a nonparametric bootstrap. In particular, cluster effects are drawn with replacement from the 100 first-stage estimates. The household errors are drawn with replacement from the first-stage residuals belonging to this specific cluster. See also Section 4.2.3.

For computing an overall state-level poverty measure, it is crucial to know at least approximately the distributions of households over municipalities in the population at the time of the recent survey: The proposed approach imputes poverty measures for the municipalities by using the dated census households. Clearly, a composite measure of those single poverty measures has to account for the number of households

[12]The proportion of all households in the 2010 census data with an income of zero amounts to 3.16 percent.

Figure 4.1: *Distributions of household residual variances and skewness in clusters*

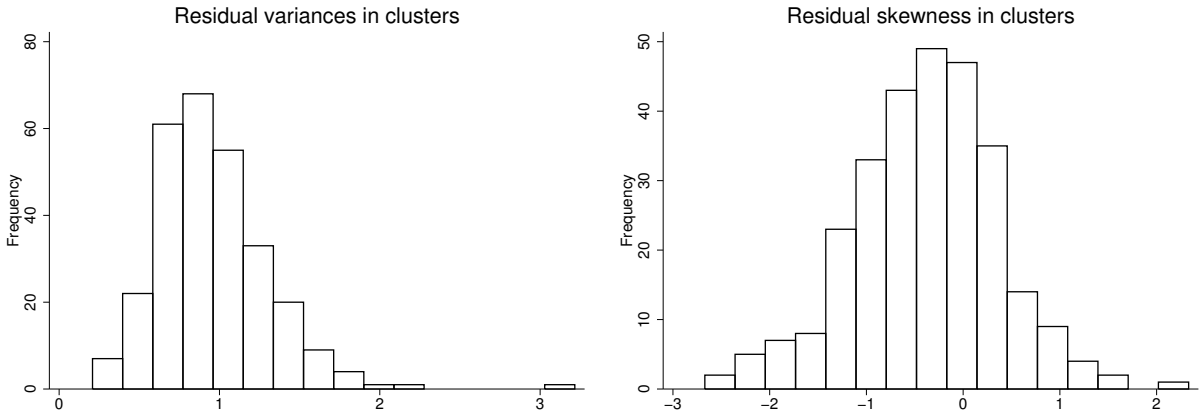



Table 4.4: *State level headcount ratio at household-level*

| | | New estimator | | | ELL estimator | | | Survey est. | |
|---|---|---|---|---|---|---|---|---|---|
| | True value | Bias | RMSE | Cov. | Bias | RMSE | Cov. | Bias | RMSE |
| $W_0(5.5)$ | 0.1076 | 0.0098 | 0.0138 | 0.8900 | 0.1020 | 0.1038 | 0.0000 | -0.0015 | 0.0137 |

in the municipalities at the time of the recent survey.[13]

We compare the performance of our estimator for the headcount ratio[14] in the state of Minas Gerais with the ELL estimator and a simple (weighted) mean of survey household incomes below the poverty line. Note that the sample is, in contrast to many real-world applications, representative and rich at the small-area level such that this purely survey-based estimator is an unbiased poverty estimator by construction. For the ELL first-stage regression, the same explanatory variables are used, yet on the household level and using the 2010 survey data. In a regression based on all households from the 2010 census, this simple model specification already yields an $R^2$ of 0.33. We conduct 300 Monte Carlo simulations with 200 bootstrap census datasets generated in each replication.

For our estimator, the coverage of the confidence intervals is below the nominal one of 95% (Table 4.4). The estimator is slightly biased which may be because of unmodeled heterogeneity in the error distribution, for example between clusters. In a regression based on all households from the 2010 census, variances and skewness of the residuals differ considerably between clusters (Figure 4.1). However, we found no clear pattern with respect to the fitted values from a first-stage regression or other explanatory variables. As the number of clusters is relatively small, already one cluster with an extreme behavior of its errors can potentially have a large effect on estimates of poverty or welfare measures. In practice, it can be challenging to detect and model such peculiarities in the error distribution. Potential remedies are discussed in Section 4.6.

Due to the bias in the headcount ratio estimator, a comparison with a (weighted) mean purely based on the survey yields a comparable, even slightly superior performance of the latter in terms of the root mean squared error. Since the distribution of the explanatory variables has changed from 2000 to 2010 (e.g., the share of households owning a phone increased from 67% to 70%), the ELL estimator is severely biased.

[13] In fact, this requirement ensures that changes in the distribution of the explanatory variables are accounted for in our approach. While it is not guaranteed to know the distribution of households at the time of the survey, it is arguably much more realistic than assuming the distribution of the explanatory variables on the household level not to change over time, as done by EEL, for instance.

[14] We also estimated the poverty gap in the same simulation setting and obtained qualitatively similar results.

Table 4.5: *State level headcount ratio on individual level*

| | | Our estimator | | | ELL estimator | | | Survey est. | |
|---|---|---|---|---|---|---|---|---|---|
| | True value | Bias | RMSE | Cov. | Bias | RMSE | Cov. | Bias | RMSE |
| $W_0(5.5)$ | 0.1259 | 0.0054 | 0.0126 | 0.9600 | 0.1249 | 0.1270 | 0.0000 | -0.0026 | 0.0179 |

So far, the poverty measures have been calculated at the household-level, while one is typically also interested in poverty measures at the individual-level such as, for instance the percentage of poor people and not households in a small area. In principle, one could conduct the first-stage regression at the individual level which is equivalent to replicating the household entries in the datasets by the respective household sizes.[15] However, when calculating an overall poverty measure from the simulated income values in the second stage, one then needs to know the number of individuals in each cluster at the time of the recent survey. The required information may be available from a previous listing exercise.

A second option starts with the first-stage regression on the household-level as described above. The smallest unit to match between the census and the survey are the municipalities. In fact, the same value of consumption is predicted on average for all households in the same municipality. For a single bootstrap simulation, they only differ by the simulated household error. Since a relationship between household size and income is assumed on the household level, typically that bigger households are poorer, one cannot randomly assign household sizes to the households. Hence, one possible remedy is to save the household sizes from the survey households and residuals from the first-stage regressions and draw them together in the bootstrap procedure in the second stage.

Another approach would impute the individual poverty measure based on its relationship with the household poverty estimators. This relationship may be hypothesized on the basis of prior knowledge or estimated from the dataset at hand. Though, if the relationship between household sizes and income differs between municipalities, these two methods do not yield unbiased state-level poverty estimators in general.

In our application, we follow the second approach, that is, we run the regression on the household level and sample residuals together with household sizes. The results indicate similar conclusions as the analyses at the household-level (Table 4.5).

## 4.6 Conclusions

In this paper we presented a new method to generate poverty maps.[16] While ours is a valid approach to combine simultaneously collected census and survey data, it also allows analysts to obtain up-to-date poverty maps when only a dated census and a more recent survey are available. In contrast to existing approaches, it has low data requirements and weak assumptions. Simulation studies showed an overall good performance. If the distribution of explanatory variables changes over time, our new estimator is superior to the most frequently used method for contemporaneous census and survey collection.

However, our approach is not immune to issues typically encountered in small area estimation techniques that combine census and survey data. In particular, variable selection and adequate modeling of apparent heteroscedasticity and differences in skewness in the residuals can be challenging. Besides, the key assumption, namely that aggregate household characteristics from the dated census relate to consumption

[15] This is due to the fact that both the household equivalent income and all explanatory variables are the same for all household members.

[16] Software code in `Stata` and `R` for the implementation of our proposed method are available on request from the authors. The recently developed `Stata` package `SAE` (Nguyen et al., 2018) can be adapted accordingly.

the same way in clusters covered by the recent survey as in clusters not covered by the recent survey, may not hold for the specific welfare estimation exercise at hand. For example, the migration pattern between census and survey collection may vary between clusters and may be correlated with the welfare status which is typically not captured by the model.

Violations of the assumptions on the error term may be partly solved by allowing for more distributional flexibility in the response variable or the error term. Rojas-Perilla et al. (2017) and the references therein provide various transformations of the response variable to achieve the validity of the assumption of identically and normally distributed error terms. A more comprehensive approach would be the application of Generalized Additive Models for Location, Scale and Shape (GAMLSS, Rigby and Stasinopoulos, 2005). This framework not only includes a huge variety of potential response distributions but also allows to link all parameters of those distributions to explanatory variables. This allows for a straightforward way to model heteroscedasticity and skewness simultaneously in one coherent model. Moreover, nonlinear and spatial effects can be integrated into the GAMLSS framework. Although model choice is also a challenging task, it might be a very interesting direction for future research to combine GAMLSS and existing small area approaches, irrespective of the time span between census and survey collection.

# 5 The (non-)significance of reporting errors in economics: Evidence from three top journals

# The (non-)significance of reporting errors in economics: Evidence from three top journals

Peter Pütz [*], Stephan B. Bruns [†]

**Abstract**

We investigate the prevalence and sources of reporting errors in hypothesis tests in three top economic journals. Reporting errors are defined as inconsistencies between reported significance levels and statistical values such as coefficients and standard errors. We analyze 30,993 tests from 370 articles and find that 34% of the articles contain at least one reporting error. Survey responses from the respective authors, replications and regression analyses suggest some simple solutions to mitigate the prevalence of reporting errors in future research. Open data and software code policies in line with a vivid replication culture seem to be most important.

**Keywords**: Reporting errors; Replications; Honest mistakes; Scientific misconduct

**JEL codes**: A11, B40, A12

[*]Chair of Statistics, Department of Economics, University of Goettingen.

[†]Corresponding author: Stephan Bruns, Center for Environmental Sciences, University of Hasselt, stephan.bruns@uni-goettingen.de.

# 5.1 Introduction

The reliability of empirical research is subject to intensive debate (e.g., Munafò et al., 2017; Wasserstein et al., 2016), with economics being no exception (e.g., Vivalt, 2019; Brodeur et al., 2018; Ioannidis et al., 2017). Most prominently, Brodeur et al. (2016) find evidence of an inflation of significant $p$-values suggesting $p$-hacking (Simonsohn et al., 2014; Leamer, 1983), HARKing (Kerr, 1998) and publication bias (Franco et al., 2014; Rosenthal, 1979) to be common practices in empirical economics, as has been shown for many other disciplines (e.g., Albarqouni et al., 2017; O'Boyle et al., 2017; Gerber and Malhotra, 2008a,b). However, reported significance levels and statistical values are usually assumed to be correct and little research has addressed the rate of errors in reported findings. In this paper, we investigate the prevalence of reporting errors in three top economic journals and shed light on potential sources. Our findings are worrying as one of three articles contains a reporting error, but remedies to reduce the prevalence of reporting errors seem to be easy to implement.

We define reporting errors as inconsistencies between reported levels of statistical significance by means of eye-catchers (mostly stars) and calculated $p$-values based on reported statistical values such as coefficients and standard errors. Errors in reporting may result from honest mistakes originating, for instance, from manually transferring empirical findings from statistical software to word processing software, updating tables in the review process, or during typesetting and insufficient proof reading by the authors. Errors may also result from scientific misconduct such as rounding down $p$-values to let them appear statistically significant (John et al., 2012). Many regression models are usually presented in one table to convince the reader of the robustness of the main findings and authors may feel tempted to add a star to one or two highlighted findings to demonstrate this robustness. Irrespective of their origin, reporting errors undermine the reliability of empirical research and future research may erroneously build on these findings (Azoulay et al., 2015).

We analyze reporting errors in 30,993 tests from 370 articles published in the *American Economic Review* (AER), *Quarterly Journal of Economics* (QJE) and *Journal of Political Economy* (JPE) by comparing reported significance levels by means of eye-catchers with calculated $p$-values based on reported statistical values such as coefficients and standard errors. We use an algorithm similar to Bruns et al. (2019) that accounts for the issue that statistical values are usually rounded and reported with low precision. This algorithm flags tests as potential reporting errors and gives authors the benefit of the doubt if rounding may be responsible for apparent reporting errors. We verify the flagged tests by contacting all authors of afflicted studies. This survey also identifies potential sources. As some flagged errors are not verified due to nonresponses by the authors, we draw a random sample of these tests and replicate the corresponding studies. Insights from the replications allow us to further verify flagged tests and to ultimately obtain a reliable estimate of the rate of reporting errors.

To the best of our knowledge, this is the first large-scale analysis of reporting errors in economics. Bruns et al. (2019) investigate the prevalence of reporting errors in 5,667 significance tests in 101 articles published in *Research Policy*, the leading outlet in innovation research at the intersection of economics and management. They detect an alarming share of 45% of articles that include at least one reporting error. Most research on reporting errors has been conducted in psychology. In a large-scale study comprising 16,695 articles and 258,105 significance tests from eight top psychology journals between 1985 and 2013, Nuijten et al. (2016) find that 49.6% of those articles are afflicted by at least one reporting error. In psychology, similar error rates are found in studies with smaller sample sizes (Bakker and Wicherts, 2014; Caperos and Pardo, 2013; Bakker and Wicherts, 2011; Wicherts et al., 2011), while Veldkamp et al. (2014) detects reporting errors in 63% of 697 investigated articles. A substantially lower share of 10.1%

is estimated by Berle and Starcevic (2007) in the field of psychiatry. In a small study comprising 44 articles published in *Nature* and the *British Medical Journal*, Garcia-Berthou and Alcaraz (2004) find a share of 31.5%. Comparing the prevalence of reporting errors at the test instead of the paper level, Bruns et al. (2019) find that 4.0% of the investigated hypothesis tests in innovation research are afflicted by a reporting error. In the field of psychology, the lowest error rates at the test level are revealed in the analyses conducted by Wicherts et al. (2011) and Bakker and Wicherts (2014) with 4.3% and 6.7%, respectively. Shares similar to or even higher than 10% are found in Nuijten et al. (2016), Veldkamp et al. (2014), Caperos and Pardo (2013), Bakker and Wicherts (2011), Berle and Starcevic (2007) and Garcia-Berthou and Alcaraz (2004).

Significance levels may be overstated, that is, eye-catchers suggest smaller $p$-values than actually true, or they may be understated, that is, eye-catchers suggest larger $p$-values than actually true. As empirical research largely focuses on rejecting null hypotheses, overstated significance levels are more consistent with incentives in academic publishing while there are little incentives to understate significance levels, except for a few articles that intend to show that the null hypothesis is true. Comparing these rates may help to reveal potential motives of reporting error in published studies. Bruns et al. (2019) find a slight imbalance towards overstated significance levels and suggestive evidence that this can be mostly attributed to authors from management rather than economics. Nuijten et al. (2016) compare the percentage of strong reporting errors which change the significance statement from significant to non-significant and *vice versa* among all reported $p$-values. They find that $p$-values reported as significant are more likely to be strong reporting errors than $p$-values reported as insignificant.

Finally, we also shed light on potential sources and predictors of reporting errors using a survey sent to authors of afflicted studies and regression analyses. There is little known about sources and predictors of reporting errors. Incorrect rounding of statistical values (Garcia-Berthou and Alcaraz, 2004; Bakker and Wicherts, 2011) and the incorrect transfer of results from statistical software to word-processing programs (Bakker and Wicherts, 2011) have been identified as sources of reporting errors. Wicherts et al. (2011) ask authors to share their research data and find that the willingness to do so is associated with a lower prevalence of reporting errors. Bruns et al. (2019) find indications that disciplines matter.

Our results show that 33.5% of the analyzed articles contain at least one reporting error corresponding to a prevalence of reporting errors of 1.3% at the test level. We find a slight imbalance towards overstated significance levels indicating that honest mistakes are likely to be the major cause of errors. Finally, many reporting errors seem to have their origin in the manual transfer of results from statistical software to word-processing software and occur when code and data are not publicly available.

The remaining part of this paper proceeds as follows: Section 5.2 introduces the dataset. Our algorithm that flags tests as potential reporting errors is described in Section 5.3. The survey is presented in Section 5.4. Section 5.5 presents replications of a large random sample of articles and the resulting estimates for the error rates in published findings. Regression analyses to identify predictors of reporting errors are shown in Section 5.6. Section 5.7 discusses all findings. Section 5.8 concludes and gives recommendations for good scientific practices to reduce reporting errors.

## 5.2 Data

Our data are based on Brodeur et al. (2016) who collected statistical values, such as coefficients with standard errors or $t$-values, for 50,078 tests from 641 articles published in the AER, the JPE and the

Table 5.1: Descriptive Statistics

| | AER | JPE | QJE | Total |
|---|---|---|---|---|
| Tests reported with coef. and se | 12247 | 4685 | 11235 | 28167 |
| Tests reported with $t/z$-statistic | 553 | 246 | 876 | 1675 |
| Tests reported with $p$-value | 447 | 66 | 638 | 1151 |

QJE between 2005 and 2011.[1] These tests relate to genuine hypothesis tests and reporting errors are particularly troublesome in these cases. Tests routinely conducted, for example for control variables or descriptive statistics, are not considered.[2]

Analyzing inconsistencies between calculated $p$-values and reported significance levels requires on the one hand sufficient statistical information to calculate a $p$-value (e.g., a coefficient with a standard error or a $t$-value) and on the other hand an eye-catcher assigning a specific level of statistical significance (e.g., stars or bold printing). These two conditions hold for 30,993 tests from 1,513 tables grouped by 370 articles. Table 5.1 presents descriptive statistics on the sample.

We extend the data by adding the reported significance level for each test. Usually significance levels are attributed by stars and the table notes clarify how the number of stars relates to different significance levels. We also added information on all significance levels used in the respective table. For example, a table may use the 0.01, 0.05, and 0.1 levels of statistical significance or the 0.001, 0.05, and 0.01 levels.

## 5.3 Flagging potential reporting errors

As a first step, we apply an algorithm to all 30,993 tests to flag potential reporting errors. This algorithm takes the low precision of reported statistical values into account and gives authors the benefit of the doubt (Bruns et al., 2019). However, in some cases tests may be falsely flagged as reporting errors as will be discussed at the end of this section. Therefore, we validate flagged tests in Section 5.4 by means of a survey among the authors and in Section 5.5 by means of replications of a random sample.

### De-rounding reported statistical values

For the majority of tests (90.9%), only coefficients and standard errors and neither test statistics nor $p$-values are reported (Table 5.1). In these cases, we calculate the ratio of coefficient and associated standard error. Generally, reconstructing the degrees of freedom used for the respective test is difficult and often impossible (Brodeur et al., 2016). We thus assume that the calculated $t$-statistic is $z$-distributed, that is standard normally distributed under the null hypothesis. This permits us to calculate the corresponding $p$-value. Of course, critical values of the $t$-distribution and standard normal distribution may differ when the degrees of freedom are small. Implications for our analysis are discussed at the end of this section.

The numbers presented in the articles are usually rounded and reported with low precision. In order to account for rounding uncertainties, we calculate intervals consistent with the reported numbers. For instance, a rounded coefficient of 0.019 and a rounded standard error of 0.010 may have their origin in non-rounded estimates from the intervals $[0.0185; 0.0194\bar{9}]$ and $[0.0095; 0.0104\bar{9}]$. The corresponding

[1] The original dataset, its description and software code in `Stata` can be downloaded from `https://www.aeaweb.org/articles?id=10.1257/app.20150044`.

[2] Brodeur et al. (2016) identify tests related to genuine hypothesis by scanning the tables and table notes and by reviewing the text where the test results are described. For further details on their reporting guidelines, see Brodeur et al. (2016) and the corresponding online appendix.

possible ratios, that we denote as $t$-values, are then given by the (rounded) interval $[1.762; 2.053]$. Using the standard normal distribution for a two-sided test, this interval is consistent with the interval $[0.0401; 0.0781]$ of possible $p$-values.[3]

When a test statistic but no $p$-value is reported (in 5.4% of the tests), the test statistic is also transformed into a $p$-value interval by taking the uncertainty due to the low precision of reported statistical values into account. When a $p$-value is reported (in 3.7% of the tests), an interval consistent with this potentially rounded $p$-value is computed. Therefore, we obtain $p$-value intervals for all tests in our dataset that are consistent with the reported statistical values.

## Diagnosis of reporting errors

Reporting errors are diagnosed if the interval of $p$-values calculated based on the reported statistical values does not overlap with the interval of $p$-values assigned by eye-catchers. If, for example, the coefficient of 0.019 with a standard error of 0.010 is labeled to be significant at the 0.05 level, a reporting error seems to be present as the ratio is 1.9 but the critical value of the 0.05 level for a two-sided test is 1.96. However, the $p$-value interval $[0.0401; 0.0781]$ that accounts for the de-rounding procedure shows that values on both sides of the threshold of 0.05 are consistent with the reported values. In these cases, no reporting error is diagnosed giving authors the benefit of the doubt. The $p$-value interval for the reported statistical significance is obtained by using information from the table notes. For example, if a table reports to use the 0.01, 0.05 and 0.1 levels, a coefficient labeled to be significant at the 0.05 level corresponds to a $p$-value interval of $[0.01; 0.05]$. If the table uses only the 0.05 and 0.1 levels, then the corresponding $p$-value interval is $[0; 0.05]$.

We refer to two different types of reporting errors: Statistical significance is defined to be overstated if the eye-catcher implies a $p$-value interval whose upper bound is smaller than the lower bound from the $p$-value interval as indicated by the reported statistical values.[4] Analogously, too large $p$-values implied by the eye-catcher are defined as understated significance level. Empirical research is mostly concerned with presenting statistically significant findings and thus overstated significance levels are usually more consistent with incentives in the research process (Brodeur et al., 2016; Ioannidis et al., 2017). Moreover, a *strong* reporting error is diagnosed if a finding turns from statistically significant to insignificant or *vice versa*. An illustration of exemplary reporting errors is given in Table 5.2.

## Prevalence of potential reporting errors

The share of articles with at least one flagged test and the share of flagged tests can be found in the first column in Table 5.3. In 50.5% (187 of 370) of the analyzed articles, our algorithm flags at least one test. At the test level, 2.1% of all tests are flagged as inconsistently reported, corresponding to 637 tests (377 understated significance levels and 260 overstated significance levels). 30% (111 of 370) of the articles are flagged as containing at least one strong reporting error corresponding to 0.69% at the test level (215 tests, among them 109 understated and 106 overstated significance levels). These initial results suggest that the prevalence of reporting errors at the article level is high while the number of affected tests is small. These results are expected as neither sloppiness nor scientific misconduct are likely to result in a large number of afflicted tests per article. While understated significance levels exceed overstated significance

[3] For one-sided tests the $p$-value interval changes accordingly. Our algorithm accounts for one-sided tests.

[4] Note that it is not known whether a reporting error originates from misreported coefficients, standard errors, test statistics or $p$-values on the one hand or significance levels on the other hand. However, the survey responses suggest that the latter is the major reason for reporting errors (see Section 5.4).

Table 5.2: Exemplary reporting errors

| ID | Coefficient | Standard Error | Lower de-rounding bound of $t$-value | Upper de-rounding bound of $t$-value | $p$-value interval as implied by reported statistical values | $p$-value interval as reported by means of eye-catchers | Type of reporting error |
|---|---|---|---|---|---|---|---|
| 1 | 0.167 | 0.128 | 1.2957 | 1.3147 | $0.1890 < p < 0.1951$ | $0.05 < p < 0.1$ | overstated |
| 2 | 0.126 | 0.067 | 1.8593 | 1.9023 | $0.0571 < p < 0.0630$ | $0.01 < p < 0.05$ | overstated |
| 3 | 0.192 | 0.115 | 1.6580 | 1.6812 | $0.0927 < p < 0.0973$ | $0.1 < p < 1$ | understated |

Notes: Coefficients and standard errors as reported in the articles. The eye-catcher and the table notes imply a $p$-value interval. The lower bound of this interval is based on the next higher threshold of statistical significance (i.e., smaller $p$-value threshold) used in the table of the respective article. For example, if the eye-catcher implies $p < 0.01$, then the lower bound is either 0.001 (if this level is indicated in the table notes) or 0 if 0.01 is the highest threshold of statistical significance used in the respective table. Analogously, if the eye-catcher does not imply statistical significance, then the lower bound is based on the lowest significance threshold used in the respective table (often 0.1) and the upper level is 1. The reported coefficients and standard errors and their de-rounding bounds imply a range of $t$-values and corresponding $p$-values. Here, we use two-sided tests and the standard normal distribution to transform the $t$-value interval into a $p$-value interval. If the lower bound of this $p$-value interval is larger than the upper bound of the $p$-value interval as implied by the eye-catchers, the reported significance level is overstated. If the upper bound of the $p$-value interval consistent with the reported statistical values is smaller than the lower bound of the $p$-value interval as implied by the eye-catchers, the reported significance level is understated. Bounds rounded to four decimal places.

levels at both the article and test level, these imbalances become less pronounced for strong reporting errors. Column two and three of Table 5.3 present refined estimates of the prevalence of reporting errors and are discussed in the next sections.

## Limitations of the algorithm

A critical step in our procedure is to treat $t$-values as standard normally distributed instead of $t$-distributed under the null hypothesis. We use this simplification as we were not able to reconstruct the degrees of freedom underlying the analyses.[5] The actual critical values from a $t$-distribution are always bigger than their analogues from the $z$-distribution, especially if the degrees of freedom are small. As a result, the number of tests flagged as errors with understated significance levels may be inflated. For example, if the $t$-statistic is equal to two and the test is labeled to be only significant at the 0.1 level but the 0.05 level is also used in the respective table, a reporting error with understated significance level seems to be present as two exceeds the critical value of the standard normal distribution for the 0.05 level (1.9600). However, the critical value of the 0.05 level for a $t$-distribution with, for example, only 50 degrees of freedom is 2.0151 and the reported significance level would actually be correct. The third example in Table 5.2 illustrates a test that may be falsely flagged as error with understated significance level due to low degrees of freedom.[6]

A second limitation of the algorithm is related to the style of reporting. The algorithm compares calculated levels of statistical significance based on reported statistical values with reported levels of statistical significance. In some cases, however, reported statistical values do not directly relate to the reported significance level. Specifically for nonlinear models marginal effects may be presented as coefficients while the reported significance levels refer to the original model. The limitations of the algorithm are addressed in the next two sections by refining the estimated rates of reporting errors based on survey responses from the authors and replications of a random sample.

[5] Sample size is often an insufficient proxy for the degrees of freedom as clustered standard errors are frequently used in the analyzed articles and `Stata` uses the number of clusters as base for the degrees of freedom.

[6] The prevalence of overstated significance levels is only affected if authors intentionally use the $z$-distribution to obtain significance levels in cases when the appropriate $t$-distribution would lead to less significant result.

Table 5.3: Prevalence of reporting errors

| | | | Flagged | Corrected by survey responses | Corrected by survey responses & replications |
|---|---|---|---|---|---|
| Article level | Any error | Overstated | 0.2811 | 0.2676 | 0.2676 |
| | | Understated | 0.3892 | 0.3351 | 0.2350 |
| | | Any | 0.5054 | 0.4568 | 0.3350 |
| | Strong error | Overstated | 0.1568 | 0.1432 | 0.1432 |
| | | Understated | 0.1892 | 0.1703 | 0.1117 |
| | | Any | 0.3000 | 0.2757 | 0.1932 |
| Test level | Any error | Overstated | 0.0084 | 0.0069 | 0.0069 |
| | | Understated | 0.0122 | 0.0099 | 0.0060 |
| | | Sum | 0.0206 | 0.0168 | 0.0129 |
| | Strong error | Overstated | 0.0034 | 0.0026 | 0.0026 |
| | | Understated | 0.0035 | 0.0030 | 0.0021 |
| | | Sum | 0.0069 | 0.0056 | 0.0047 |

Notes: Prevalence of any and strong reporting errors at article and test level. "Overstated" means overstated significance level, "Understated" means understated significance level. At the article level, the share of articles with at least one overstated and understated significance level and any of them is given. The estimates are calculated based on our algorithm to flag tests in the raw data (first column), after taking into account the survey responses (second column) and after additionally including the information from the replication task.

## 5.4 Survey

We sent a survey via email to all authors whose articles contain at least one flagged test to validate the findings of our algorithm and to shed light on the sources of reporting errors. The authors were provided with the statistical values of the flagged tests.

### Survey questions

In our first question, the authors were asked where the reporting error occurred, that is, whether it occurred in the coefficient, standard error, test statistic, $p$-value or eye-catcher. Two further response options were "I don't know" and "There is no reporting error". The second question concerned the sources of the potential reporting error. As possible response options, we offered: "Error occurred while transferring results from statistical software to word processing software such as Word or Latex", "Error occurred while updating tables during the research/review process", "Error occurred in typesetting by the publisher and remained undetected in proofreading", "Reporting error is falsely diagnosed due to low degrees of freedom of the corresponding test (algorithm to detect reporting errors relies on critical values of the standard normal distribution)", "I don't know", "Other reason" and "If 'other reason' applies, please specify".[7] We sent one reminder to nonresponding authors after three weeks and waited further three weeks before stopping the data collection. We promised the authors to treat their answers anonymously.[8]

[7]The authors sometimes did not fill in the survey attached to the email but replied to our questions directly via e-mail. In these cases, we translated their answers to the survey.

[8]The email and an exemplary survey can be found in an online appendix at `http://www.stephanbbruns.de/reperrors`.

Table 5.4: Where is the reporting error? ($n = 303$)

| Coefficient | Stand. error | Test statistic | $p$-value | Eye-catcher | There is no error | I don't know |
|---|---|---|---|---|---|---|
| 3 | 12 | 0 | 0 | 121 | 133 | 34 |
| (1.0%) | (4.0%) | (0.0%) | (0.0%) | (39.9%) | (43.9%) | (11.2%) |

Table 5.5: Why is there a reporting error? ($n = 170$)

| Transfer | Updating | Typesetting | I don't know | Other reason |
|---|---|---|---|---|
| 58 | 15 | 15 | 50 | 32 |
| (34.1%) | (8.8%) | (8.8%) | (29.4%) | (18.8%) |

Notes: "Transfer" refers to the incorrect transfer of results from statistical software to word processing software such as *Word* or *LaTeX*. "Updating" indicates that an error occurred while updating tables during the research/review process. "Typesetting" means that an error occurred in typesetting by the publisher and remained undetected in proofreading.

## Responses

The survey was responded by 88 of 164 contacted authors (53.7%) with regard to 98 articles (52.4% of all articles containing at least one flagged test) and 309 flagged tests (48.5% of all flagged tests).[9] Authors replied that 133 or 43.9% of all flagged tests are no reporting errors (Table 5.4). Most of the remaining 170 flagged tests were confirmed to be errors in the eye-catchers.

Among these 170 reporting errors, the incorrect transfer of results from statistical software to word processing software such as *Word* or *LaTeX* ("transfer") was the main explanation for reporting errors (Table 5.5).[10] This answer was given for 34.1% of the errors, three times more often than each of the two other main sources: Table updating during the research / review process ("updating") and typesetting by the publisher ("typesetting"). Almost 30% of the errors were not explained. Other sources were given for about 18% of the errors.[11]

## Classification of flagged tests

We classified a flagged test as reporting error if an error was confirmed by the authors, that is, if they replied that the error occurred at a specific place (e.g., coefficient) or due to a particular reason (e.g., typesetting). We cross-checked the 133 flagged tests which the authors replied to be no reporting errors. As can be seen in Table 5.6, in 21.8% of the cases the authors plausibly argued that low degrees of freedom caused the test to be falsely flagged ("low df"). In other instances, the same reason was given, but we were not able to confirm the argumentation. Most importantly, errors with overstated significance levels cannot be falsely flagged due to low degrees of freedom. We classified those answers as wrong and did the same for other implausible or illogical answers.[12]

A further reason for falsely flagged tests by our algorithm were deviations from the common reporting

[9] Six of these flagged tests were due to a misalignment, namely wrong formatting in one article. However, the reported statistical values and eye-catchers were consistent if the obvious misalignment was accounted for. Besides, the author pointed out that an erratum was published. Therefore, we classified the flagged tests as no errors and treat the remaining 303 flagged tests as benchmark in the calculations.

[10] One author replied that he transmitted the `Stata` results to his co-author via phone, who then entered the numbers into a word processing program.

[11] These include answers which were not possible to assign reasonably to the other response categories as, for example, rounding errors, and meaningless answers such as that software did not report significance levels for the respective table.

[12] For example, some authors argued that they interpreted significance levels as less than or equal to some value instead of strictly less. However, the probability to obtain a $p$-value exactly equal to a threshold is zero and it is more likely that in fact a rounding error or another type of error occurred. One author argued that the 0.01 level of significance implies the 0.05 level, but he used the 0.01, 0.05, and 0.1 levels in the same table for other estimates.

Table 5.6: Why is there no reporting error? ($n = 133$)

| Coder's fault | Nonstandard reporting | Low df | Low df possible | Wrong answer |
|---|---|---|---|---|
| 3 | 45 | 29 | 32 | 24 |
| (2.3%) | (33.8%) | (21.8%) | (24.1%) | (18.0%) |

Notes: "Coder's fault" refers to a error in the original coding or by us. "Nonstandard reporting" means that the reporting style deviates from the common one used for OLS regressions and thus leads to a flagged test which is no reporting error, though. "Low df" stands for low degrees of freedom which cause a falsely flagged test since our algorithm to detect reporting errors relies on critical values of the standard normal distribution. "Low df possible" means that the authors did not give a reason why there is no reporting error, but we found that low degrees of freedom are a likely reason that there is indeed no reporting error. "Wrong answer" indicates that the reason of the author why there should not be a reporting error is implausible.

style which is used for OLS regressions. For example, if a probit model was used, authors sometimes reported the coefficients and standard errors of marginal effects, but the eye-catchers referred to the significance test corresponding to the original probit coefficient. Such a reporting style, which we call nonstandard, was the reason for 45 or 33.8% falsely flagged tests stemming from five articles with one article accounting for 26 of these tests. Although the answers were plausible to us after validation, a distinct explanation of the reporting style is missing in four of the five articles.

If the authors argued that there was no reporting error but without reasoning, we examined whether data was erroneously coded, low degrees of freedom, or a nonstandard reporting style could have been the reason for falsely flagged tests. We found that for 32 cases low degrees of freedom are a possible explanation and agreed with the authors' responses ("Low df possible"). Data was falsely coded for three flagged tests.

## Update of error rates

In sum, 109 of the initially flagged tests are likely to be no errors with the main reasons of low degrees of freedom (61) and nonstandard reporting style (45), see Table 5.6. Of the 260 tests initially flagged as errors with overstated significance levels, 33.1% were confirmed to be indeed errors, 16.2% were falsely flagged as errors, and 50.8% remain without verification from the authors either because the authors did not reply to the survey or replied "I do not know" to both survey questions, see Tables 5.4 and 5.5.[13] The 16.2% of tests that were falsely flagged correspond to 42 tests of which 40 used a nonstandard reporting style and two were incorrectly coded. As becomes evident in column two of Table 5.3, the rate of overstated significance levels decreases at the test level moderately from 0.84% to 0.69% for all errors and 0.34% to 0.26% for strong errors while the prevalence at the article level decreases only slightly from 28.1% to 26.8% for all errors and 15.7% to 14.3% for strong errors. The error rate at the article level remains similar as only a few articles account for many falsely flagged errors due to nonstandard reporting.

Of the 377 tests initially flagged as understated significance levels, 20.7% were confirmed to be indeed errors, 17.8% were falsely flagged as errors, and 61.5% remain without verification. The 17.8% tests that were falsely flagged correspond to 67 tests of which 61 were flagged because of low degrees of freedom, five due to a nonstandard reporting style and one due to a coding error. We expect the number of falsely flagged errors to be higher for understated significance levels due to the limitations of the algorithm. The error rate at the test level moderately reduces from 1.22% to 0.99% for all errors and 0.35% to 0.30% for strong errors while at the article level the prevalence decreases from 38.9% to 33.5% for all errors and

[13]The authors sometimes gave a reason via mail or additional comment in the survey why they cannot replicate their results and explain why the tests were flagged. The main reason was that they did not have access to the software code anymore.

18.9% to 17.0% (Table 5.3, column two). Again, reduction at the article level is smaller as articles often have multiple flagged errors of which not all result from low degrees of freedom.

## 5.5 Replications

The survey shed light on 279 (43.8%) of the flagged tests. 60.9% of them were correctly flagged as reporting errors and 39.1% were incorrectly flagged as reporting errors. The survey leaves 358 (56.2%) of the flagged tests without manual verification by the authors. Out of these, 91.6% are due to no response from the authors and 8.4% due to the authors' reply "I do not know" to both survey questions. In the following, we estimate how many of the flagged tests without manual verification by the authors are indeed reporting errors by replicating afflicted studies.

### Replication strategy

We took a random sample of 30% from all flagged tests resulting in 119 tests from 64 articles. As we tried to replicate all flagged tests of these 64 articles the sample comprises 83.2% of the flagged tests without verification by the authors (298 flagged tests). We searched the web for data and software code for the respective articles and used `Stata` 12.1 and `R` 3.5.1 (Windows) to conduct the replications.

If we were able to replicate the reported statistical values of the flagged test exactly, we checked whether the $p$-value obtained in the replication is consistent with the $p$-value interval reported in the article by means of an eye-catcher. In this case, we classified the flagged test as no reporting error. However, sometimes the replicated statistical values are similar but not identical to those reported in the article. In these cases, we used the values reported in the article and calculated the corresponding $p$-value by using the procedure given by the authors' code (degrees of freedom and distribution under the null hypothesis). If this $p$-value was consistent with the reported eye-catcher, we again classified the flagged tests as no reporting errors. This procedure allows us to give the benefit of the doubt to the authors in case software was updated and, for example, the same command produces slightly different standard errors today. Replication success in this sense is the ability to judge whether a reporting error is present without necessarily replicating the exact statistical values of the original article. If data and software code were available but the replication results differed substantially from the original findings, we classified the test as not replicable but not as a reporting error. In this study, we define reporting errors as inconsistencies between reported levels of statistical significance and calculated $p$-values based on reported statistical information. Of course, non-replicable results may also be considered as a type of reporting error.

### Findings

In 179 or 60.1% of the 298 flagged tests, we were not able to replicate the results. These tests belong to 50 out of 64 articles (78.1%), among them six articles containing both replicable and non-replicable tests. The main reason for non-replication was that data or software code was not provided. This was the case for 150 tests (83.8% of 179) from 40 articles (80.0% of 50).[14] In 23 cases (12.8%) from 8 articles (16.0%), we were not able to obtain similar estimates compared to those of the published article, despite data and `Stata` or `R` code were available. For the remaining six cases (3.4%) from two articles (4.1%) we did not replicate the tests since a different software was used.[15]

[14] Data confidentiality was the reason for seven tests from two papers why no data was provided.
[15] The detailed replication results can be found online at `http://www.stephanbbruns.de/reperrors`.

We managed to (partially) replicate 20 studies comprising 119 out of the 298 flagged tests (29 overstated and 90 understated). As we tried to replicate all flagged tests of the sampled articles, the following estimates at the test and article level are weighted accordingly. We found that all of the flagged overstated significance levels were indeed reporting errors, that is, the estimated rate of correctly flagged overstated significance levels is 100%. For the understated significance levels, the corresponding rate is 45.8%. The vast majority of falsely flagged understated significance levels was due to low degrees of freedom.[16] Regarding strong reporting errors, the estimated rates of correctly flagged strong errors were 100% for overstated significance levels and 48.9% for understated significance levels.

To estimate the rate of articles with at least one correctly flagged (strong) error, we followed the same strategy. More specifically, we divided the number of articles with at least one flagged test of a particular kind (overstated, understated or any error) after the replication by its analog before the replications were conducted. The resulting rates of articles with at least one correctly flagged error are 100% (overstated), 49.9% (understated) and 51.2% (any error). Note that all articles with incorrectly flagged understated significance levels had no other unverified flagged test. The shares of articles with at least one correctly flagged strong error were estimated to be 100% (overstated), 39.8% (understated) and 45.5% (any error). These estimates are fairly similar to the ones at the test level as presented above but based on smaller sample sizes of replicated articles: There were only seven articles with at least one overstated significance level, 19 with at least one understated significance level and 20 with at least one flagged test of any kind before the replications were conducted. For articles with strong reporting errors, these numbers reduce to four, eleven and thirteen.

[16] In one case, a nonstandard reporting style caused the flagging by our algorithm.

### Update of error rates

The sum of the flagged overstated and understated significance levels which were neither verified by the authors nor replicated was multiplied by the respective shares of correctly detected errors as calculated above. The same was done for the sum of articles with at least one of such non-verified tests, for overstated and understated significance levels and any flagged test, respectively. Following this strategy, we find that in 33.5% of the investigated articles, there is at least one reporting error (Table 5.3, column 3). On the test level, 1.3% of all tests are afflicted by a reporting error. For strong reporting errors these numbers reduce to 19.3% and 0.5% (Table 5.3, column 3). Overall, the rates of overstated significance levels exceed slightly those of understated significance levels.[17]

## 5.6 Exploratory regression analyses

In addition to the survey responses, we explore potential predictors of reporting errors applying logistic regression models. The dependent variable indicates whether an article includes at least one (strong) reporting error or not. We implement the corrections obtained from the survey responses and replications. Table 5.10 and Table 5.11 show that the distribution of (strong) errors over articles is heavily skewed. We run logistic regressions at the article level to avoid the high influence of outliers on the estimates that may occur in an analysis at the test level. Since we did not specify hypotheses beforehand, our analyses should be deemed purely exploratory.

### Model specification

The explanatory variables are taken from the large set of variables gathered by Brodeur et al. (2016) and we focus on those that vary at the article level. In particular, we include the journal, the rough research field, whether negative results are put forward, whether a theoretical model is used, data availability, code availability, the year of publication, the authors' average years since their PhDs as well as the numbers of authors, research assistants and individuals thanked, tables and tests, and the shares of editors and tenured authors among the authors as predictors. More details on the variables and descriptive statistics are given in the appendix and in Tables 5.8 and 5.9. We dropped variables that are essentially equivalent to those we used. For example, the share of authors who are editors or members of editorial boards at the time of publication is very similar to the same share prior to the publication year. Likewise, we did not include variables containing categories with very few observations that are difficult to group as, for instance, one variable indicating the specific field of an article that has field categories which apply to only one article. We reran the models 500 times using a nonparametric bootstrap.

### Results

The results in the first column and third column of Table 5.7 are very similar, that is, the probabilities to observe an article with at least one reporting error and an article with at least one strong reporting error can be explained by the same variables. Histograms for the estimates across bootstrap samples are presented in Figure 5.1 and Figure 5.2. In most of the 500 bootstrap samples, articles without theoretical models, from the field of macroeconomics in comparison to microeconomics and with more tests are more

[17] Robustness checks for the estimated error rates can be found in the appendix.

Table 5.7: Regression results

| | Any error | Any error | Strong error | Strong error |
|---|---|---|---|---|
| Intercept | 72.405 | 61.442 | 85.992 | 64.730 |
| | [-116.909; 275.045] | [-129.843; 261.180] | [-127.602; 325.400] | [-163.490; 303.102] |
| Year | -0.036 | -0.031 | -0.043 | -0.032 |
| | [-0.137; 0.058] | [-0.130; 0.065] | [-0.163; 0.063] | [-0.151; 0.082] |
| Journal of Political Economy | 0.603 | 0.136 | 0.774 | 0.099 |
| | [0.079; 1.208] | [-0.528; 0.838] | [0.168; 1.470] | [-0.706; 0.937] |
| Quarterly Journal of Economics | -0.060 | -1.006 | -0.105 | -1.426 |
| | [-0.544; 0.498] | [-1.913; -0.233] | [-0.712; 0.533] | [-2.584; -0.452] |
| Field: Macroeconomics | 0.677 | 0.693 | 0.552 | 0.597 |
| | [0.194; 1.161] | [0.212; 1.198] | [-0.014; 1.091] | [0.069; 1.162] |
| No. of authors | -0.145 | -0.156 | -0.097 | -0.110 |
| | [-0.411; 0.073] | [-0.428; 0.061] | [-0.429; 0.161] | [-0.458; 0.172] |
| Share of editors among authors | -0.383 | -0.429 | -0.472 | -0.544 |
| | [-1.054; 0.300] | [-1.079; 0.238] | [-1.226; 0.232] | [-1.289; 0.174] |
| Share of tenured authors | 0.636 | 0.818 | 0.365 | 0.626 |
| | [-0.191; 1.554] | [0.017; 1.725] | [-0.567; 1.326] | [-0.265; 1.688] |
| Authors' average years since PhD | -0.002 | -0.014 | 0.005 | -0.012 |
| | [-0.047; 0.042] | [-0.059; 0.029] | [-0.041; 0.053] | [-0.059; 0.032] |
| No. of research assistants thanked | -0.044 | -0.041 | -0.041 | -0.039 |
| | [-0.119; 0.011] | [-0.118; 0.016] | [-0.156; 0.028] | [-0.155; 0.031] |
| No. of individuals thanked | 0.014 | 0.013 | 0.014 | 0.014 |
| | [-0.013; 0.042] | [-0.015; 0.043] | [-0.016; 0.046] | [-0.017; 0.048] |
| Negative results put forward | -0.217 | -0.218 | -0.143 | -0.140 |
| | [-0.767; 0.326] | [-0.800; 0.334] | [-0.833; 0.529] | [-0.893; 0.570] |
| With theoretical model | -0.455 | -0.440 | -0.560 | -0.566 |
| | [-0.972; -0.030] | [-0.979; -0.008] | [-1.208; -0.097] | [-1.238; -0.115] |
| No. of tables | 0.081 | 0.098 | -0.041 | -0.017 |
| | [-0.020; 0.191] | [-0.005; 0.209] | [-0.163; 0.051] | [-0.146; 0.084] |
| No. of tests | 0.005 | 0.005 | 0.006 | 0.007 |
| | [0.002; 0.009] | [0.002; 0.008] | [0.004; 0.011] | [0.004; 0.012] |
| Data available | | 0.023 | | -0.150 |
| | | [-0.657; 0.702] | | [-1.015; 0.670] |
| Code available | | -1.042 | | -1.305 |
| | | [-1.971; -0.239] | | [-2.431; -0.360] |
| $n$ | 367 | 367 | 367 | 367 |
| Pseudo $R^2$ | 0.094 | 0.081 | 0.077 | 0.102 |

Notes: Results from logistic regressions. The dependent variable is whether an article contains at least one reporting error or not (first two columns) or at least one strong reporting error (last two columns). Lower and upper bounds of 90% bias corrected and accelerated (BCa) intervals based on 500 bootstrap replicates in brackets.

likely to include at least one (strong) reporting error.[18] Likewise, articles in the JPE seem to be afflicted by at least one (strong) reporting error more frequently than in the AER and QJE. One of the reasons might be the journal policy which determines whether data and software code have to be published. In our regression sample, in none of the articles in the QJE data or code are available on the website of the journal. The articles in the AER are mostly accompanied by data (82.6%) and code (92.1%), while for the JPE data and code are available in 46.7% and 48.3% of the articles, respectively. If the two variables data and code availability are added to the regression, both the articles in JPE and the AER are associated more often than the QJE with at least one (strong) reporting error over most of the bootstrap replicates (second and fourth column in Table 5.7). To put it differently, the AER performs comparably better if one does not control for the two transparency variables. Apparently, the stringent data and code requirements of the AER lower the probability to find articles with errors in their published issues. The described effects are associated with considerable uncertainty though. Moreover, the effects of most explanatory variables under consideration vary substantially from negative to positive values over 90% of the bootstrap samples. The explanatory power of the models is limited as can be seen by the low Pseudo $R^2$.

[18] The reference categories for nominal variables are given in Table 5.8.

## 5.7 Discussion

### Prevalence of reporting errors

Our analyses show that reporting errors in economics are a relevant issue but suggest that the rate of errors is smaller in comparison to other fields. The error rates of 1.3% at the test level and 33.5% at the article level are substantially lower than those found in similar investigations from other academic disciplines. At the test level, it is by far the lowest one observed in comparable studies in the literature so far. While Bruns et al. (2019), Bakker and Wicherts (2014) and Wicherts et al. (2011) find error rates between 4.0% and 6.7%, the rates in most studies are similar to or even higher than 10% (Nuijten et al., 2016; Veldkamp et al., 2014; Caperos and Pardo, 2013; Bakker and Wicherts, 2011; Berle and Starcevic, 2007; Garcia-Berthou and Alcaraz, 2004). Similarly, higher error rates between 45.1% and 63% are found at the article level in all studies in the field of psychology (Nuijten et al., 2016; Bakker and Wicherts, 2014; Veldkamp et al., 2014; Caperos and Pardo, 2013; Bakker and Wicherts, 2011; Wicherts et al., 2011) and in the field of innovation research by Bruns et al. (2019).

Clearly, comparisons between different studies are difficult. For instance, Nuijten et al. (2016) analyze crawled statistical values reported in the text, excluding tables. They detect a higher prevalence of reporting errors in a sample which was also manually checked for reporting errors by Wicherts et al. (2011). In our procedure, we are rather conservative because of the de-rounding procedure. In some cases, we obtain large $p$-value intervals based on the reported statistical values which are consistent with more than one level of statistical significance.

As the prevalence of reporting errors at the test level is substantially lower in our study compared to all other studies which are mainly from psychology, our findings still indicate that a difference between academic disciplines exists. This is also in line with the results from Bruns et al. (2019) who use a similar algorithm as we do to flag reporting errors in innovation research. Their estimated prevalence of reporting errors seems to be in between the one found by us for top economic journals and those found by others in (top) psychology journals. At the article level, comparisons between fields are even more disputable as the numbers of tests per article may differ substantially. Nevertheless, it is interesting to see that articles in our study are less likely to contain at least one reporting error than articles in most studies from other fields.

We find overstated significance levels to be slightly overrepresented compared to understated significance levels. Understated significance levels are usually neither favourable nor consistent with the incentive system in academic publishing. Overstated significance levels have, moreover, a higher probability to survive the review process as additional stars are less likely to be caught by the authors than missing stars (Nuijten et al., 2016). As the imbalance towards overstated significance levels in our study is small, the vast majority of reporting errors are likely to result from honest mistakes. Our findings are consistent with Bruns et al. (2019) and Nuijten et al. (2016) who also find a slight imbalance towards overstated significance levels.

Since our dataset includes only articles from three very prestigious journals in economics, it is not clear whether our results hold for economics in general. One might argue that less prestigious journals are less thorough in the review process and thus more likely to be afflicted by reporting errors.

Our replication sample that we use to estimate the error rates is nonrandom as we only replicate those flagged tests for which we found data and software code in `R` or `Stata`. Our replication based estimate is that 100% of the flagged tests that imply an overstated significance level are correct. We explore

the robustness of this finding using a variable gathered by Brodeur et al. (2016) indicating the type of statistical model/test used to cross-check whether we missed tests with a nonstandard reporting style among the flagged tests. A nonstandard reporting style was identified by author responses in 45 cases. In 41 of these 45 cases, this variable shows that a logit, probit or a model different to linear regression was applied. Among the flagged tests from which we drew the replication sample, the same holds for only twelve overstated significance levels. Of these twelve flagged tests, seven were indeed reporting errors as shown by the replication and five were not replicated. This suggests that the estimated share of the correctly flagged tests with overstated significance levels is accurate.

### Sources and predictors of reporting errors

The survey responses suggest that the manual transfer of results from statistical software to word processing software is a major source of reporting errors. Even though automatic procedures had existed a long time before 2005 when our time frame of investigated articles starts, for example *outreg* for `Stata` (Gallup, 1998, 1999, 2000), it is conceivable that a manual transfer might still have been common practice in those days and it might still be today. Copy-paste as a source of reporting error is also identified by Bakker and Wicherts (2011).

Our regression analyses shows that the prediction of reporting errors is difficult with the help of the available variables. One reason is measurement error in the dependent variable as our algorithm to detect reporting errors is not perfect. Articles with theoretical contribution as well as articles from microeconomics have a lower probability to contain at least one (strong) reporting error. Possible interpretations, that we deem rather speculative, are left to the reader. More apparent are differences between journals. Based on our findings and those from Wicherts et al. (2011) there is at least weak evidence for a lower prevalence of reporting errors if data and software codes are shared. This is a very conclusive finding and can be embedded into the general call for more open data and code policies (e.g, Chang and Li, 2018). They succeed to qualitatively replicate the key results of 32.8% of 67 articles in the field of macroeconomics without contacting the authors. We consider a nonrandom sample from microeconomics and macroeconomics and only try to replicate the flagged tests in articles without manual verification by the authors. Therefore, our study is hardly comparable to Chang and Li (2018), but two general similarities can be identified. First, we also find a very low rate of successful replications as we were only able to replicate all flagged tests in 21.9% of 64 articles. Second, the main reason for a failed replication both in Chang and Li (2018) and our study is missing public data and software code.

## 5.8 Conclusions and recommendations

By investigating more than 30,000 hypotheses tests of central hypotheses from 370 articles published in three top economic journals, we find that 1.3% percent of all tests are afflicted by a reporting error. Strong reporting errors that make a significant finding non-significant or *vice versa* occur in 0.5% of all tests. This relates to 33.5% of the articles being afflicted by at least one reporting error while 19.3% are afflicted by at least one strong reporting error. Although these error rates are low in comparison to those found in related literature from other research fields, it is worrying that one of three articles contains at least one reporting error, even in top economics journals. We also find a slight imbalance towards overstated significance levels. Since understated significance levels are usually neither favourable nor consistent with the incentive system in academic publishing, our findings indicate that the vast majority of reporting errors are likely to result from honest mistakes.

Our analysis suggests three approaches to reduce the rate of reporting errors in future research:

First, data and code should be made available by the researchers. Our findings indicate that availability of data and code may reduce the probability of an article to contain a reporting error. Authors that provide data and code are likely to carefully check whether the uploaded code indeed replicates the published tables and this is likely to reduce the probability of a reporting error. More importantly, transparency facilitates replications and permits others to check the accuracy of the published findings. Replications can be incentivized by introducing replication sections in top journals (Coffman et al., 2017) and by accepting positive replications for publication (Mueller-Langer et al., 2017).

Second, eye-catchers to denote statistical significance should be banned. According to the authors who participated in our survey, most of the detected reporting errors stem from errors in the eye-catchers. Moreover, it is conceivable that eye-catchers distract authors and reviewers from studying intensively the actual statistical values and checking them for consistency. The *American Economic Association* indeed nowadays forbids authors to use stars to denote statistical significance in their journals, among others in the AER.[19] However, eye-catchers are still omnipresent in other journals such as the JPE and the QJE. Eye-catchers sustain the bad practice to judge scientific results in a binary way according to arbitrary thresholds. Instead, reporting effect sizes and confidence intervals facilitates the interpretation of economic significance rather than statistical significance (Cumming, 2014). Moreover, classical Bayesian methods (Kruschke and Liddell, 2018) or reverse Bayesian techniques (Colquhoun, 2017) may also be fruitful. In any case, one should never consider one study in an isolated manner but include all evidence available (Amrhein et al., 2018; McShane et al., 2018; Greenland et al., 2016).

Third, errors can be reduced in the research, review and publication process by simple measures. Automatic software should be used to transfer statistical results to word processing software. Algorithms such as the one used in this paper could be used in the review process and by the authors themselves after the typesetting as suggested by Nuijten et al. (2016).

[19] `https://www.aeaweb.org/journals/aer/submissions/accepted-articles/styleguide`

# 5.A Appendix

## Descriptive statistics

In the following, we provide additional information on some of the variables used in the regression analyses in Section 5.6 in the order as depicted in Tables 5.8 and 5.9.

- Negative results: Whether the article presents at least one null result in contrast to only positive results as contribution. As there are only seven articles which only put forward null results and 53 which report positive and null results, we grouped these two response categories together.
- Theoretical model: Whether the article includes a theoretical contribution.
- Data availability: Whether the data of the article is available on the respective website of the journal.
- Code availability: Whether the software code of the article is available on the respective website of the journal.
- Share of editors among the authors: Share of authors who are member of an editorial board at the time of publication.
- Share of tenured authors: Share of authors who are full professors at the time of publication.
- Authors' average years since PhD: Authors' average years since PhD at the time of publication.
- Number of individuals thanked: Number of individuals thanked excluding research assistant and referees.
- Number of tables: Number of tables reporting results of hypothesis tests that test central hypotheses, tables with several panels are treated as one table.
- Number of tests: Number of hypothesis tests that test central hypotheses.

Details on the original reporting guidelines for the variables can be found in the online appendix of Brodeur et al. (2016).

## Robustness checks

In the following, we describe two robustness checks to estimate the prevalence of reporting errors. We account for the corrections by the survey and replication, so that the results should be compared to column three of Table 5.3.

A potential issue regarding these estimates became evident when we investigated the last decimals of all reported statistical values: The share of zeros was only about 5.6%, whereas each of the other digits comprised 9.8-11.2% of the cases. A manual check proved that Brodeur et al. (2016) occasionally dropped zeros as last decimals. To test whether this increased uncertainty in the rounding procedure leads to a substantially lower rate of detected reporting errors, we reran our analyses after dropping the last decimals if they are equal to zero. We found almost no difference in the results, suggesting that dropped zeros at the end of the statistical values do not contribute notably to underestimated rates of reporting errors (Table 5.12, column one).

As a further robustness check, we examined whether potentially trimmed decimals, that is, a reported coefficient of 1.4 was originally 1.48, for example, affect the results. While emphasizing that this would still constitute erroneous rounding by the authors, namely rounding up when rounding down would be adequate, the estimated rates change only marginally (Table 5.12, column two).

Table 5.8: Distribution of discrete variables over tests and articles

| | Number of Articles | Number of Tests |
|---|---|---|
| American Economic Review | 180 | 13247 |
| Journal of Political Economy | 61 | 4997 |
| Quarterly Journal of Economics | 129 | 12749 |
| Macroeconomics | 85 | 8142 |
| Microeconomics | 285 | 22851 |
| Negative results: yes | 60 | 5848 |
| Negative results: no | 310 | 25145 |
| With theoretical model | 109 | 8964 |
| Without theoretical model | 261 | 22029 |
| Data available | 177 | 13553 |
| Data not available | 193 | 17440 |
| Code available | 195 | 15214 |
| Code not available | 175 | 15779 |

Table 5.9: Continuous variables on article level

| Variable | n | Min | Q1 | Median | Mean | Q3 | Max | SD |
|---|---|---|---|---|---|---|---|---|
| Year | 370 | 1 | 3.0 | 4.4 | 5.0 | 6.0 | 7 | 2.0 |
| Number of authors | 370 | 1 | 2.0 | 2.2 | 2.0 | 3.0 | 5 | 0.9 |
| Share of editors among authors | 370 | 0 | 0.0 | 0.4 | 0.3 | 0.6 | 1 | 0.4 |
| Share of tenured authors | 370 | 0 | 0.0 | 0.3 | 0.3 | 0.5 | 1 | 0.3 |
| Authors' average years since PhD | 367 | -2 | 5.5 | 10.1 | 9.0 | 12.8 | 41 | 6.6 |
| Number of research assistants thanked | 370 | 0 | 0.0 | 2.2 | 1.0 | 3.0 | 27 | 3.5 |
| Number of individuals thanked | 370 | 0 | 6.0 | 11.2 | 9.5 | 15.0 | 45 | 7.7 |
| Number of tables | 370 | 1 | 2.0 | 4.1 | 4.0 | 5.0 | 15 | 2.4 |
| Number of tests | 370 | 2 | 30.2 | 83.8 | 60.0 | 107.8 | 587 | 82.9 |

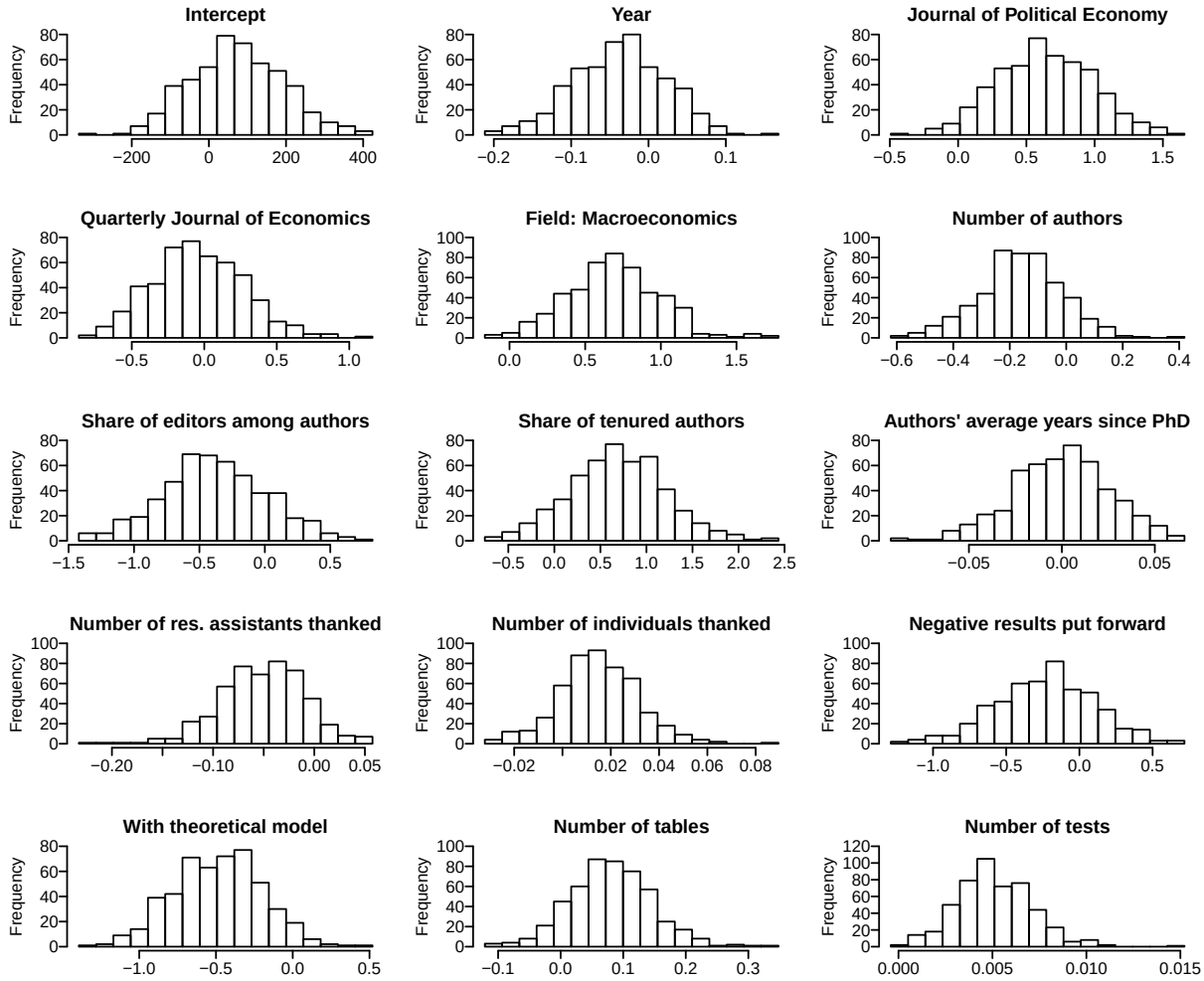
Table 5.10: Distribution of errors over articles

| Number of errors | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 16 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 208 | 67 | 31 | 23 | 16 | 7 | 4 | 2 | 1 | 3 | 3 | 2 | 1 | 1 | 1 |

Table 5.11: Distribution of strong errors over articles

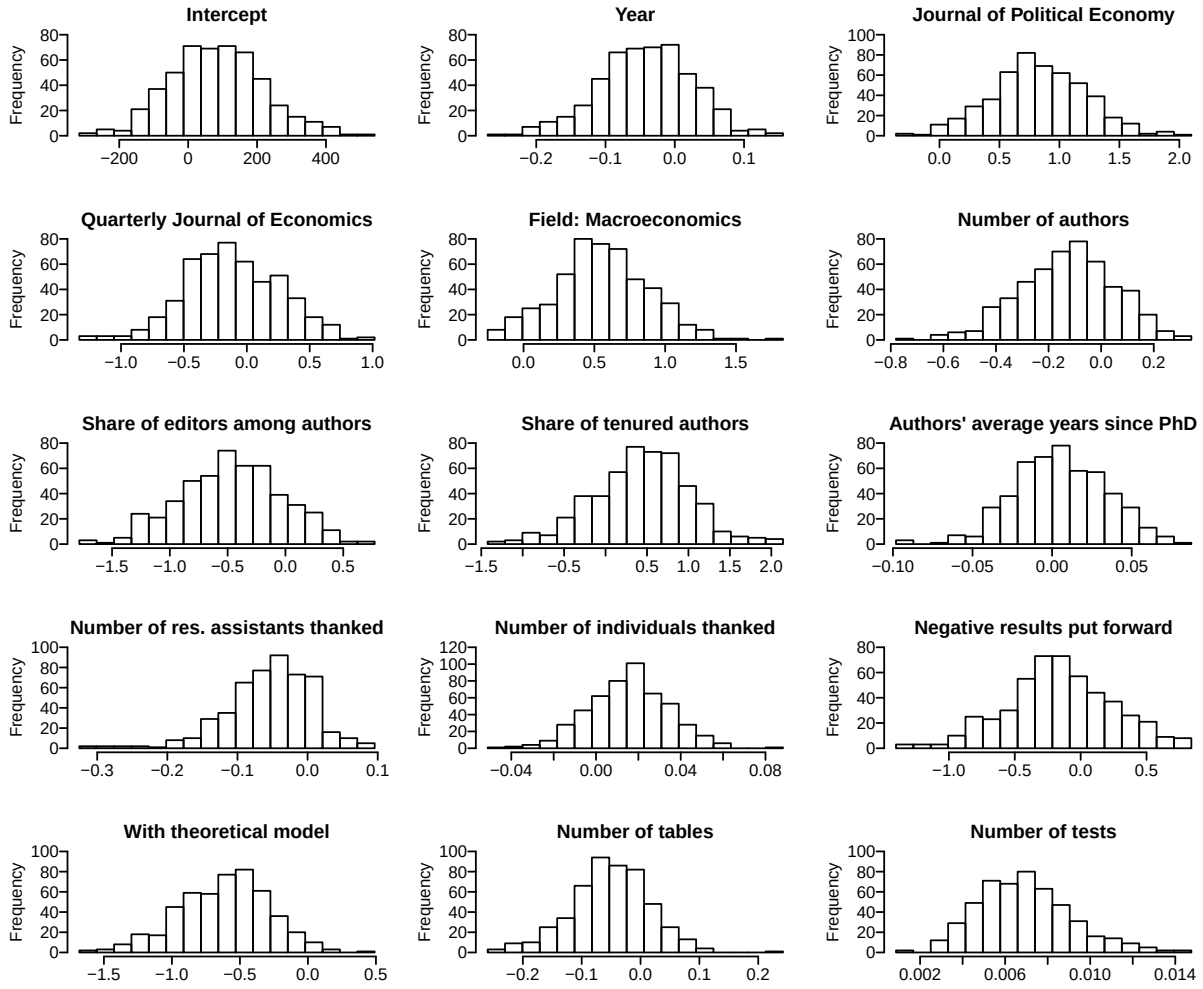
| Number of errors | 0 | 1 | 2 | 3 | 4 | 5 | 15 |
|---|---|---|---|---|---|---|---|
| Frequency | 274 | 61 | 21 | 7 | 4 | 2 | 1 |

Figure 5.1: Distribution of bootstrap estimates 1




Notes: The histogram shows the distribution of 500 bootstrap estimates for all explanatory variables in a logistic regression with the dependent variable indicating whether an article contains at least one reporting error or not.

Figure 5.2: Distribution of bootstrap estimates 2




Notes: The histogram shows the distribution of 500 bootstrap estimates for all explanatory variables in a logistic regression with the dependent variable indicating whether an article contains at least one strong reporting error or not.

Table 5.12: Prevalence of reporting errors - robustness checks

| | | | Zeros removed | Trimmed decimals |
|---|---|---|---|---|
| Article level | Any error | Overstated | 0.2622 | 0.2595 |
| | | Understated | 0.2337 | 0.2202 |
| | | Any | 0.3295 | 0.3173 |
| | Strong error | Overstated | 0.1405 | 0.1378 |
| | | Understated | 0.1117 | 0.1042 |
| | | Any | 0.1932 | 0.1799 |
| Test level | Any error | Overstated | 0.0067 | 0.0065 |
| | | Understated | 0.0059 | 0.0054 |
| | | Sum | 0.0125 | 0.0119 |
| | Strong error | Overstated | 0.0026 | 0.0025 |
| | | Understated | 0.0021 | 0.0019 |
| | | Sum | 0.0047 | 0.0044 |

Notes: Prevalence of any and strong reporting errors at test and article level. "Overstated" means overstated significance level, "Understated" means understated significance level. At the article level, the share of articles with at least one overstated and understated significance level and any of them is given. The estimates are calculated after including information the survey responses and the replication exercise. The first column shows error rates calculated under the condition that the last decimals of the reported statistical values are removed if they are equal to zero. The second column shows error rates calculated under the condition that authors trimmed the reported statistical values instead of rounding properly.

# 6 Conclusions and outlook

This thesis presents advanced models and practices in applied statistics with regard to causality, prediction, and the replicability of results. In the following, the papers' contributions are grouped according to these three domains. Corresponding future research developments are sketched from a general viewpoint, while suggestions for specific future research directions were given in the four papers in the previous chapters. Some general remarks close this thesis.

## 6.1 Causality

The papers in Chapters 2 and 3 are both methodological contributions to regression models frequently applied to identify causal effects. The first paper in Chapter 2 relaxes assumptions on the shape of the relationship between explanatory variables and the dependent variable. Instead of determining specific functional shapes of the covariate effects a priori, these are allowed to be data-driven and potentially nonlinear by using penalized splines. The main contribution of the paper in Chapter 2 is to use such splines for fixed effects panel data models. These models are often applied to control for unobserved time-invariant heterogeneity by including fixed effects which are allowed to be correlated with other covariates. A first-difference approach for the inclusion of penalized splines in such a framework is presented. Furthermore, a corresponding fast way of inference via simultaneous confidence bands is provided.

The main novelty of the second paper in Chapter 3 lies in the combination of methods to estimate causal relationships with GAMLSS. This combination allows to model causal effects not only on the conditional expectation but on all aspects of the distribution of the dependent variable. Using data from the famous Progresa program, it is shown step-by-step how the proposed methodology provides insights which go unnoticed in common regression frameworks.

As more flexible models may be more meaningful and appropriate for specific research questions and data situations, future research should promote their use to the scientific community. In particular, the ongoing focus on linear regression specification and changes in the expectation of the dependent variable should be enhanced by more advanced methods as presented in this thesis. In this regard, methods to establish causality may not be as "harmless" as Angrist and Pischke (2008) suggest. Thus, step-by-step guides tailored to the specific research fields may help to make these advanced methods accessible to scientists. They may also be useful to exchange the knowledge from scientific fields. For example, relaxing the functional form of covariate effects by using penalized splines is frequently done in ecology but rarely observed in economics.

Another interesting research direction considers how to exploit increasing data availability and computational power when estimating causal effects. Machine learning techniques such as boosting, lasso, random forests and or deep neural networks were originally dedicated to prediction tasks. Applying them to research questions concerning causality seems to be a promising avenue (e.g., Chernozhukov et al., 2018; Wager and Athey, 2018; Belloni et al., 2017, 2018).

## 6.2 Prediction

The paper in Chapter 4 is a methodological contribution to a small area prediction problem. In particular, it is shown how to estimate reliable up-to-date poverty maps when an outdated census and a more recent survey are available. The proposed approach has lower data requirements and weaker assumptions than commonly applied poverty mapping methods that are tailored to situations with simultaneous survey and census collection periods. Although the focus of the paper is on poverty mapping, the methodology presented is applicable to a wide range of outcome measures and research questions.

Extensions for this and basically all prediction tasks include more distributional flexibility in the response variable, nonlinear covariate effects as well as variable and model selection. Mayr et al. (2012) provide an appropriate overarching approach by boosting GAMLSS. While GAMLSS allow for a wide range of response distributions and flexible covariate specification, boosting serves as a regularization and variable selection tool. A fully nonparametric deep learning approach which estimates the whole conditional density of the dependent variable given a potentially high-dimensional predictor might be also fruitful (Li et al., 2019).

In general, machine learning methods are powerful and increasingly applied in many scientific prediction tasks and also in causal analyses as discussed in Section 6.1. However, as every approach to analyze data, they include numerous pitfalls (Domingos, 2012). As they often deviate from classical statistical modeling approaches with which applied researchers are familiar, their application in specific research fields needs further elaboration in the form of guides as, for instance, provided by Mullainathan and Spiess (2017) for economists.

## 6.3 Replicability

The paper in Chapter 5 considers replicability from two perspectives. On the one hand, the paper shows that one third of 370 articles published in top economic journals would not be entirely replicable even when data and code were available because they contain at least one reporting error. On the other hand, systematic replications demonstrate that data and code are available for the minority of 64 articles under consideration and even data and code availability do not guarantee a successful replication of the study results. Proposed solutions to enhance replicability comprise the ban of eye-catcher, the automatic check of the consistency of reported statistical information before publication, incentives for a more vivid replication culture as well as open data and code policies.

The paper in Chapter 3 includes a replication of a study. While using the same data but a different statistical model, very similar results are obtained when looking at the expectation of the response distribution. The paper thus directly contributes to a more vivid replication culture that several scientists call for (e.g., Chang and Li, 2018; Clemens, 2017; Coffman et al., 2017).

With regard to reporting errors, studies in all research fields except from psychology (e.g., Nuijten et al., 2016) are rare. Therefore, future studies are needed and could investigate and compare the drivers and prevalence of reporting errors in different research fields.

Replicability in a wider sense implies the replication of statistical results in a new study. The definition of replication success in this sense is controversial as described in Section 1.3. It has to be further discussed when and how replication success can be measured in a reasonable way and how comparable and reliable studies can be implemented. In any case, additional issues arise when replicating a study

using a new dataset. Research practices such as $p$-hacking (e.g., Simonsohn et al., 2014) and HARKing (Hypothesis After Research Results are Known, Kerr, 1998) lead to selective reporting of statistical results. The consequence is not only unexpectedly low replicability, measured by whatever method, but also a distorted body of evidence for the cumulative research process. Future studies may monitor and evaluate to what extent solutions to low replicability (or reproducibility) as proposed by Munafò et al. (2017) are applied in single research fields.

## 6.4 Some closing remarks

The focus of this thesis lies on advanced models and practices in three domains of applied statistics mainly in the context of regression models with applications in economics. Regression models are applied in many research fields such as biology, epidemiology, medicine, physics, psychology, and sociology and the arguments presented in this thesis carry over to these fields. More generally, challenges in finding an adequate model in consideration of the research question and data at hand are inherent to any statistical modeling approach. Likewise, ensuring the replicability of statistical results is an indispensable task irrespective of the type of statistical analysis used.

Applied statistics is naturally confronted with more concerns within and outside the three domains covered in this thesis, as the highly debated domain of statistical inference shows (e.g., Wasserstein et al., 2016). Accordingly, this thesis is just a further component in the large building of applied statistics.

With due care and modesty, I nonetheless hope that this thesis contributes to scientific communities that apply statistical methods. After all, contemporary and future researchers have a larger toolkit to choose from than their predecessors. They can address specific data situations and research questions more appropriately. And hopefully, they will apply the proposed instruments to enhance the replicability of their statistical results.

# Bibliography


Agostini, C. A., Brown, P. H. and Roman, A. C.: 2010, Poverty and inequality among ethnic groups in Chile, *World Development* **38**(7), 1036–1046.

Albarqouni, L. N., López-López, J. A. and Higgins, J. P.: 2017, Indirect evidence of reporting biases was found in a survey of medical research studies, *Journal of Clinical Epidemiology* **83**, 57–64.

Amemiya, T.: 1974, The nonlinear two-stage least-squares estimator, *Journal of Econometrics* **2**(2), 105–110.

Amemiya, T.: 1977, The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model, *Econometrica* **45**(4), 955.

Amrhein, V., Trafimow, D. and Greenland, S.: 2018, Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication, *The American Statistician* (forthcoming).

Angelucci, M. and De Giorgi, G.: 2009, Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?, *American Economic Review* **99**(1), 486–508.

Angrist, J., Chernozhukov, V. and Fernández-Val, I.: 2006, Quantile regression under misspecification, with an application to the U.S. wage structure, *Econometrica* **74**(2), 539–563.

Angrist, J. D. and Pischke, J.-S.: 2008, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Angrist, J. D. and Pischke, J.-S.: 2010, The credibility revolution in empirical economics: How better research design is taking the con out of econometrics, *Journal of economic perspectives* **24**(2), 3–30.

Antoniadis, A., Gijbels, I. and Verhasselt, A.: 2012, Variable selection in additive models using P-splines, *Technometrics* **54**(4), 425–438.

Araujo, M. C., Ferreira, F. H., Lanjouw, P. and Özler, B.: 2008, Local inequality and project choice: Theory and evidence from Ecuador, *Journal of Public Economics* **92**(5-6), 1022–1046.

Azoulay, P., Furman, J. L., Krieger, J. L. and Murray, F.: 2015, Retractions, *Review of Economics and Statistics* **97**(5), 1118–1136.

Baker, M.: 2016, 1,500 scientists lift the lid on reproducibility, *Nature* (7604), 452–454.

Bakker, M. and Wicherts, J. M.: 2011, The (mis)reporting of statistical results in psychology journals, *Behavior Research Methods* **43**(3), 666–678.

Bakker, M. and Wicherts, J. M.: 2014, Outlier removal and the relation with reporting errors and quality of psychological research, *PLoS ONE* **9**(7), 1–9.

Baltagi, B. H. and Li, D.: 2002, Series estimation of partially linear panel data models with fixed effects, *Annals of Economics and Finance* **3**(1995), 103–116.

Bärnighausen, T., Røttingen, J.-A., Rockers, P., Shemilt, I. and Tugwell, P.: 2017, Quasi-experimental study designs series–paper 1: Introduction: Two historical lineages, *Journal of clinical epidemiology* **89**, 4–11.

Bassett, G. and Koenker, R.: 1982, An empirical quantile function for linear models with iid errors, *Journal of the American Statistical Association* **77**(378), 407–415.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S. and Umlauf, N.: 2015, BayesX - Software for inference in structured additive regression models. Version 3.0.2. Available from `http://www.bayesx.org`.

Bello-Orgaz, G., Jung, J. J. and Camacho, D.: 2016, Social big data: Recent achievements and new challenges, *Information Fusion* **28**, 45–59.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C.: 2017, Program evaluation and causal inference with high-dimensional data, *Econometrica* **85**(1), 233–298.

Belloni, A., Chernozhukov, V. and Kato, K.: 2018, Valid post-selection inference in high-dimensional approximately sparse quantile regression models, *Journal of the American Statistical Association* (forthcoming).

Berger, J. O. and Berry, D. A.: 1988, Statistical analysis and the illusion of objectivity, *American Scientist* **76**(2), 159–165.

Berle, D. and Starcevic, V.: 2007, Inconsistencies between reported test statistics and p-values in two psychiatry journals, *International Journal of Methods in Psychiatric Research* **16**(4), 202–207.

Betti, G., Dabalen, A., Ferré, C. and Neri, L.: 2013, Updating poverty maps between censuses: A case study of Albania, in *Poverty and Exclusion in the Western Balkans*, Springer, pp. 55–70.

Bitler, M. P., Gelbach, J. B. and Hoynes, H. W.: 2017, Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? Evidence from a social experiment, *The Review of Economics and Statistics* **99**(4), 683–697.

Blundell, R. W. and Powell, J. L.: 2004, Endogeneity in semiparametric binary response models, *The Review of Economic Studies* **71**(3), 655–679.

Bondell, H. D., Reich, B. J. and Wang, H.: 2010, Noncrossing quantile regression curve estimation, *Biometrika* **97**(4), 825–838.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L. and Bärnighausen, T.: 2014, Regression Discontinuity Designs in Epidemiology: Causal Inference without Randomized Trials, *Epidemiology* **25**(5), 729–737.

Brodeur, A., Cook, N. and Heyes, A. G.: 2018, Methods matter: P-hacking and causal inference in economics. IZA discussion paper 11796, Institute for the Study of Labor, Bonn, available from `https://www.econstor.eu/bitstream/10419/185256/1/dp11796.pdf`.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y.: 2016, Star wars: The empirics strike back, *American Economic Journal: Applied Econometrics* **8**(1), 1–32.

Bruns, S. B., Asanov, I., Bode, R., Dunger, M., Funk, C., Hassan, S. M., Hauschildt, J., Heinisch, D., Kempa, K., König, J., Lips, J., Verbeck, M., Wolfschütz, E. and Buenstorf, G.: 2019, Errors and biases in reported significance levels: Evidence from innovation research, *Research Policy* **48**(9).

Bui, T. D. and Nguyen, C. V.: 2017, Spatial poverty reduction in Vietnam: An application of small area estimation, *Economics Bulletin* **37**(3), 1785–1796.

Calonico, S., Cattaneo, M. D., Farrell, M. H. and Titiunik, R.: 2018, Regression discontinuity designs using covariates, *The Review of Economics and Statistics* (forthcoming).

Calonico, S., Cattaneo, M. D. and Titiunik, R.: 2014, Robust nonparametric confidence intervals for regression-discontinuity designs, *Econometrica* **82**(6), 2295–2326.

Cameron, A. C. and Miller, D. L.: 2015, A practitioner's guide to cluster-robust inference, *Journal of Human Resources* **50**(2), 317–372.

Cameron, D. B., Mishra, A. and Brown, A. N.: 2016, The growth of impact evaluation for international development: How much have we learned?, *Journal of Development Effectiveness* **8**(1), 1–21.

Caperos, J. M. and Pardo, A.: 2013, Consistency errors in p-values reported in Spanish psychology journals, *Psicothema* **25**(3), 408–414.

Casey, K., Glennerster, R. and Miguel, E.: 2012, Reshaping institutions: Evidence on aid impacts using a preanalysis plan, *The Quarterly Journal of Economics* **127**(4), 1755–1812.

Celidoni, M.: 2013, Vulnerability to poverty: An empirical comparison of alternative measures, *Applied Economics* **45**(12), 1493–1506.

Chang, A. C. and Li, P.: 2018, Is economics research replicable? Sixty published papers from thirteen journals say "often not", *Critical Finance Review* **7**.

Chaudhuri, S., Jalan, J. and Suryahadi, A.: 2002, Assessing household vulnerability to poverty from cross-sectional data: A Methodology and estimates from Indonesia. Discussion paper 0102-52, department of economics, Columbia University, available from `https://academiccommons.columbia.edu/doi/10.7916/D85149GF`.

Chavez-Martin del Campo, J.: 2006, Does conditionality generate heterogeneity and regressivity in program impacts? The Progresa experience. Working Paper 127042, Department of Applied Economics and Management, Cornell University, available from `https://econpapers.repec.org/paper/agscudawp/127042.htm`.

Chen, J., Gao, J. and Li, D.: 2013, Estimation in partially linear single-index panel data models with fixed effects, *Journal of Business & Economic Statistics* **31**(3), 315–330.

Chen, J., Li, D. and Gao, J.: 2013, Non- and semi-parametric panel data models: A selective review. Working paper 18/13, Monash University, department of econometrics and business statistics, available from `https://www.monash.edu/business/ebs/research/publications/ebs/wp18-13.pdf`.

Chernick, M. R., González-Manteiga, W., Crujeiras, R. M. and Barrios, E. B.: 2011, *Bootstrap Methods*, Springer.

Chernozhukov, V., Fernández-Val, I. and Melly, B.: 2013, Inference on counterfactual distributions, *Econometrica* **81**(6), 2205–2268.

Chernozhukov, V., Fernández-Val, I. and Weidner, M.: 2018, Network and panel quantile effects via distribution regression. Arxiv working paper 1803.08154, available from `https://arxiv.org/pdf/1803.08154.pdf`.

Chernozhukov, V., Fernández-Val, I. and Galichon, A.: 2010, Quantile and probability curves without crossing, *Econometrica* **78**(3), 1093–1125.

Chernozhukov, V. and Hansen, C.: 2006, Instrumental quantile regression inference for structural and treatment effect models, *Journal of Econometrics* **132**(2), 491–525.

Claeskens, G., Krivobokova, T. and Opsomer, J. D.: 2009, Asymptotic properties of penalized spline estimators, *Biometrika* **96**(3), 529 –544.

Claeskens, G. and Van Keilegom, I.: 2003, Bootstrap confidence bands for regression curves and their derivatives, *The Annals of Statistics* **31**(6), 1852–1884.

Clemens, M. A.: 2017, The meaning of failed replications: A review and proposal, *Journal of Economic Surveys* **31**(1), 326–342.

Coffman, L. C., Niederle, M. and Wilson, A. J.: 2017, A proposal to organize and promote replications, *American Economic Review* **107**(5), 41–45.

Colquhoun, D.: 2017, The reproducibility of research and the misinterpretation of p-values, *Royal Society Open Science* **4**(12), 171085.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A. and Goodner, B.: 2007, Spatially adaptive bayesian penalized splines with heteroscedastic errors, *Journal of Computational and Graphical Statistics* **16**(2), 265–288.

Cumming, G.: 2014, The new statistics: Why and how, *Psychological Science* **25**(1), 7–29.

Cumming, G. and Maillardet, R.: 2006, Confidence intervals and replication: Where will the next mean fall?, *Psychological methods* **11**(3), 217.

Das, S. and Chambers, R.: 2017, Robust mean-squared error estimation for poverty estimates based on the method of Elbers, Lanjouw and Lanjouw, *Journal of the Royal Statistical Society. Series A: Statistics in Society* **180**(4), 1137–1161.

De Boor, C.: 2001, *A Practical Guide to Splines*, Springer.

Deaton, A. and Kozel, V.: 2005, Data and dogma: The great Indian poverty debate, *World Bank Research Observer* **20**(2), 177–199.

Demombynes, G. and Özler, B.: 2005, Crime and local inequality in South Africa, *Journal of Development Economics* **76**(2), 265–292.

Dette, H. and Volgushev, S.: 2008, Non-crossing non-parametric estimates of quantile curves, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(3), 609–627.

Djebbari, H. and Smith, J.: 2008, Heterogeneous impacts in Progresa, *Journal of Econometrics* **145**(1), 64–80.

Domingos, P. M.: 2012, A few useful things to know about machine learning, *Commun. acm* **55**(10), 78–87.

Dunn, P. K. and Smyth, G. K.: 1996, Randomized quantile residuals, *Journal of Computational and Graphical Statistics* **5**(3), 236–244.

Efron, B.: 1987, Better bootstrap confidence intervals, *Journal of the American Statistical Association* **82**(397), 171–185.

Efron, B. and Tibshirani, R. J.: 1994, *An Introduction to the Bootstrap*, CRC press.

Eilers, P. H. C. and Marx, B. D.: 1996, Flexible Smoothing with B-Splines and Penalties, *Statistical Science* **11**(2), 89–102.

Einav, L. and Levin, J.: 2014, Economics in the age of big data, *Science* **346**(6210), 1243089/1–1243089/5.

Elbers, C., Fujii, T., Lanjouw, P., Özler, B. and Yin, W.: 2007, Poverty alleviation through geographic targeting: How much does disaggregation help?, *Journal of Development Economics* **83**(1), 198–213.

Elbers, C., Lanjouw, J. O. and Lanjouw, P.: 2003, Micro-level estimation of poverty and inequality, *Econometrica* **71**(1), 355–364.

Elbers, C. and van der Weide, R.: 2014, Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. World Bank policy research working paper 6962, The World Bank, available from `http://documents.worldbank.org/curated/en/712781468338974024/pdf/WPS6962.pdf`.

Emwanu, T., Hoogeveen, J. G. and Okiira Okwi, P.: 2006, Updating poverty maps with panel data, *World Development* **34**(12), 2076–2088.

Erevelles, S., Fukawa, N. and Swayne, L.: 2016, Big data consumer analytics and the transformation of marketing, *Journal of Business Research* **69**(2), 897–904.

Eubank, A. R. L. and Speckman, P. L.: 1993, Confidence bands in nonparametric regression, *Journal of the American Statistical Association* **88**(424), 1287–1301.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B.: 2013, *Regression: Models, Methods and Applications*, Springer.

Ferrer-i Carbonell, A. and Frijters, P.: 2004, How important is methodology for the estimates of the determinants of happiness?, *Economic Journal* **114**(497), 641–659.

Firpo, S.: 2007, Efficient semiparametric estimation of quantile treatment effects, *Econometrica* **75**(1), 259–276.

Firpo, S., Fortin, N. M. and Lemieux, T.: 2009, Unconditional quantile regressions, *Econometrica* **77**(3), 953–973.

Foresi, S. and Peracchi, F.: 1995, The conditional distribution of excess returns: An empirical analysis, *Journal of the American Statistical Association* **90**(430), 451.

Foster, J., Greer, J. and Thorbecke, E.: 1984, A class of decomposable poverty measures, *Econometrica:* **52**(3), 761–766.

Franco, A., Malhotra, N. and Simonovits, G.: 2014, Publication bias in the social sciences: Unlocking the file drawer, *Science* **345**(6203), 1502–1595.

Frandsen, B. R., Frölich, M. and Melly, B.: 2012, Quantile treatment effects in the regression discontinuity design, *Journal of Econometrics* **168**(2), 382–395.

Frijters, P. and Beatton, T.: 2012, The mystery of the u-shaped relationship between happiness and age, *Journal of Economic Behavior & Organization* **82**(2-3), 525–542.

Frölich, M. and Melly, B.: 2013, Unconditional quantile treatment effects under endogeneity, *Journal of Business & Economic Statistics* **31**(3), 346–357.

Gallup, J. L.: 1998, Formatting regression output for published tables, *Stata Technical Bulletin* **8**(46), 28–30.

Gallup, J. L.: 1999, Revision of outreg, *Stata Technical Bulletin* **49**(23), 170–171.

Gallup, J. L.: 2000, Revision of outreg, *Stata Technical Bulletin* **9**(49), 9–13.

Garcia-Berthou, E. and Alcaraz, C.: 2004, Incongruence between test statistics and p-values in medical papers, *BMC Medical Research Methodology* **4**(1), 13.

Geraci, A., Fabbri, D. and Monfardini, C.: 2018, Testing exogeneity of multinomial regressors in count data models: Does two-stage residual inclusion work?, *Journal of Econometric Methods* **7**(1), 1–19.

Gerber, A. and Malhotra, N.: 2008a, Do statistical reporting standards affect what is published? Publication bias in two leading political science journals, *Quarterly Journal of Political Science* **3**(3), 313–326.

Gerber, A. S. and Malhotra, N.: 2008b, Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?, *Sociological Methods & Research* **37**(1), 3–30.

Gibson, J.: 2018, Forest loss and economic inequality in the Solomon Islands: Using small-area estimation to link environmental change to welfare outcomes, *Ecological Economics* **148**, 66–76.

Goodman, S. N., Fanelli, D. and Ioannidis, J. P.: 2016, What does research reproducibility mean?, *Science Translational Medicine* **8**(341), 341ps12.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. and Altman, D. G.: 2016, Statistical tests, p-values, confidence intervals, and power: A guide to misinterpretations, *European Journal of Epidemiology* **31**(4), 337–350.

Guadarrama, M., Molina, I. and Rao, J. N. K.: 2016, A comparison of small area estimation methods for poverty mapping, *Statistics in Transition New Series* **1**(17), 41–66.

Hajargasht, G.: 2009, Nonparametric panel data models: A penalized spline approach. CEPA working paper WP05/2009, University of Queensland, School of Economics, available from `https://sites.google.com/site/ghgasht/NonpPanelSpline.pdf?attredirects=0`.

Hamermesh, D. S.: 2013, Six decades of top economics publishing: Who and how?, *Journal of Economic Literature* **51**(1), 162–72.

Härdle, W., Huet, S., Mammen, E. and Sperlich, S.: 2004, Bootstrap inference in semiparametric generalized additive models, *Econometric Theory* **20**(2), 265–300.

Haslett, S. J.: 2016, Small area estimation using both survey and census unit record data, in *Analysis of Poverty Data by Small Area Estimation*, John Wiley & Sons, pp. 325–348.

Haslett, S. J., Isidro, M. C. and Jones, G.: 2010, Comparison of survey regression techniques in the context of small area estimation of poverty, *Survey Methodology* **36**(2), 157–170.

Hastie, T., Tibshirani, R. and Friedman, J.: 2009, *The Elements of Statistical Learning: Data Mining, Inference and Prediction (second edition)*, Springer.

He, X.: 1997, Quantile curves without crossing, *The American Statistician* **51**(2), 186–192.

Healy, A. J., Jitsuchon, S. and Vajaragupta, Y.: 2003, Spatially disaggregated estimates of poverty and inequality in Thailand. Working Paper, Massachusetts Institute of Technology and Thailand Development Research Institute, available from `http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Healy.DisaggregatedThailand.pdf`.

Heckman, J. J.: 1978, Dummy endogenous variables in a simultaneous equation system, *Econometrica* **46**(4), 931–59.

Henderson, D. J., Carroll, R. J. and Li, Q.: 2008, Nonparametric estimation and testing of fixed effects panel data models, *Journal of Econometrics* **144**(1), 257–275.

Hofner, B., Mayr, A. and Schmid, M.: 2016, gamboostLSS: An R package for model building and variable selection in the GAMLSS framework, *Journal of Statistical Software* **74**(1), 1–31.

Imbens, G. and Kalyanamaran, K.: 2012, Optimal bandwidth choice for the regression discontinuity estimator, *Review of Economic Studies* **79**(3), 933–959.

Imbens, G. W. and Lemieux, T.: 2008, Regression discontinuity designs : A guide to practice, *Journal of Econometrics* **142**(2), 615–635.

Ioannidis, J. P., Stanley, T. D. and Doucouliagos, H.: 2017, The power of bias in economics research, *Economic Journal* **127**(605), F236–F265.

Isidro, M. C.: 2010, Intercensal updating of small area estimates. PhD thesis, Massey University, New Zealand, available from `https://mro.massey.ac.nz/bitstream/handle/10179/2053/02_whole.pdf`.

Isidro, M. C., Haslett, S. and Jones, G.: 2016, Extended structure preserving estimation (ESPREE) for updating small area estimates of poverty, *Annals of Applied Statistics* **10**(1), 451–476.

John, L. K., Loewenstein, G. and Prelec, D.: 2012, Measuring the prevalence of questionable research practices with incentives for truth telling, *Psychological Science* **23**(5), 524–532.

Kauermann, G., Krivobokova, T. and Fahrmeir, L.: 2009, Some asymptotic results on generalized penalized spline smoothing, *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **71**(2), 487–503.

Kerr, N. L.: 1998, Harking: Hypothesizing after the results are known, *Personality and Social Psychology Review* **2**(3), 196–217.

Kijima, Y. and Lanjouw, P.: 2003, Poverty in India during the 1990s: A regional perspective. Policy research working paper 3141, The World Bank, available from `http://documents.worldbank.org/curated/en/938731468771597743/129529322_20041117183102/additional/multi0page.pdf`.

Kleiber, C. and Kotz, S.: 2003, *Statistical Size Distributions in Economics and Actuarial Sciences*, John Wiley & Sons.

Klein, N., Kneib, T., Klasen, S. and Lang, S.: 2015, Bayesian structured additive distributional regression for multivariate responses, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **64**(4), 569–591.

Klein, N., Kneib, T. and Lang, S.: 2015, Bayesian generalized additive models for location, scale, and shape for zero-inflated and overdispersed count data, *Journal of the American Statistical Association* **110**(509), 405–419.

Klein, N., Kneib, T., Lang, S. and Sohn, A.: 2015, Bayesian structured additive distributional regression with an application to regional income inequality in Germany, *Annals of Applied Statistics* **9**(2), 1024–1052.

Kline, P. and Santos, A.: 2012, A score based approach to wild bootstrap inference, *Journal of Econometric Methods* **1**(1), 23–41.

Kneib, T.: 2013, Beyond Mean Regression, *Statistical Modelling* **13**(4), 275–303.

Koenker, R.: 2005, *Quantile Regression*, Cambridge University Press.

Koenker, R. and Bassett, G.: 1978, Regression quantiles, *Econometrica* **46**(1), 33–50.

Krivobokova, T. and Kauermann, G.: 2007, A note on penalized spline smoothing with correlated errors, *Journal of the American Statistical Association* **102**(480), 1328–1337.

Krivobokova, T., Kneib, T. and Claeskens, G.: 2010, Simultaneous Confidence Bands for Penalized Spline Estimators, *Journal of the American Statistical Association* **105**(490), 852–863.

Kruschke, J. K. and Liddell, T. M.: 2018, The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective, *Psychonomic Bulletin and Review* **25**(1), 178–206.

Laporte, A. and Windmeijer, F.: 2005, Estimation of panel data models with binary indicators when treatment effects are not constant over time, *Economics Letters* **88**(3), 389–396.

Leamer, E. E.: 1983, Let's take the con out of econometrics, *The American Economic Review* **73**(1), 31–43.

Lee, D. S. and Lemieux, T.: 2010, Regression discontinuity designs in economics, *Journal of Economic Literature* **48**(2), 281–355.

Li, G., Peng, H. and Tong, T.: 2013, Simultaneous confidence band for nonparametric fixed effects panel data models, *Economics Letters* **119**(3), 229–232.

Li, R., Bondell, H. D. and Reich, B.: 2019, Deep distribution regression. Arxiv working paper 1903.06023, available from `https://arxiv.org/abs/1903.06023`.

Loader, C. R. and Sun, J.: 1997, Robustness of tube formula based confidence bands, *Journal of Computational and Graphical Statistics* **6**(2), 242–250.

López Ulloa, B. F., Møller, V. and Sousa-Poza, A.: 2013, How does subjective well-being evolve with age? A literature review, *Journal of Population Ageing* **6**(3), 227–246.

Lucas, R. E.: 2007, Adaptation and the set-point model of subjective well-being: Does happiness change after major life events?, *Current Directions in Psychological Science* **16**(2), 75–79.

Ludwig, J. and Miller, D. L.: 2007, Does head start improve children's life chances? Evidence from a regression discontinuity design, *The Quarterly Journal of Economics* **122**(1), 159–208.

Machado, J. A. F. and Mata, J.: 2005, Counterfactual decomposition of changes in wage distributions using quantile regression, *Journal of Applied Econometrics* **20**(4), 445–465.

Mammen, E., Støve, B. and Tjøstheim, D.: 2009, Nonparametric additive models for panels of time series, *Econometric Theory* **25**(2), 442–481.

Marhuenda, Y., Molina, I., Morales, D. and Rao, J. N.: 2017, Poverty mapping in small areas under a twofold nested error regression model, *Journal of the Royal Statistical Society. Series A: Statistics in Society* **180**(4), 1111–1136.

Marra, G. and Radice, R.: 2011, A flexible instrumental variable approach, *Statistical Modelling* **11**(6), 581–603.

Mayr, A., Fenske, N., Hofner, B. and Kneib, T.: 2012, Generalized additive models for location, scale and shape for high dimensional data - a flexible approach based on boosting, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**(3), 403–427.

McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J.: 2018, Abandon statistical significance, *The American Statistician* (forthcoming).

Meager, R.: 2016, Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature. Working paper, Massachusetts Institute of Technology, available from `http://economics.mit.edu/files/12292`.

Melly, B.: 2005, Decomposition of differences in distribution using quantile regression, *Labour Economics* **12**(4), 577–590.

Minnesota Population Center: 2017, Integrated Public Use Microdata Series, International: Version 6.5.

Molina, I. and Rao, J. N.: 2010, Small area estimation of poverty indicators, *Canadian Journal of Statistics* **38**(3), 369–385.

Mueller-Langer, F., Fecher, B., Harhoff, D. and Wagner, G. G.: 2017, The economics of replication. IZA Discussion Paper 10533, Institute for the Study of Labor, Bonn, available from `https://www.econstor.eu/bitstream/10419/161156/1/dp10533.pdf`.

Mullainathan, S. and Spiess, J.: 2017, Machine learning: An applied econometric approach, *Journal of Economic Perspectives* **31**(2), 87–106.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie Du Sert, N., Simonsohn, U., Wagenmakers, E. J., Ware, J. J. and Ioannidis, J. P.: 2017, A manifesto for reproducible science, *Nature Human Behaviour* **1**(1), 1–9.

Mundlak, Y.: 1978, On the pooling of time series and cross section data, *Econometrica: journal of the Econometric Society* **46**(1), 69–85.

National Academies of Sciences, E., Medicine et al.: 2016, *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*, National Academies Press.

Neumann, M. H. and Polzehl, J.: 1998, Simultaneous bootstrap confidence bands in nonparametric regression, *Journal of Nonparametric Statistics* **9**(4), 307–333.

Nguyen, M. C., Corral, P., Azevedo, J. P. and Zhao, Q.: 2018, sae: A stata package for unit level small area estimation. Policy research working paper 8630, The World Bank, available from `http://documents.worldbank.org/curated/en/398721540906483895/pdf/WPS8630.pdf`.

Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S. and Wicherts, J. M.: 2016, The prevalence of statistical reporting errors in psychology (1985–2013), *Behavior Research Methods* **48**(4), 1205–1226.

O'Boyle, E. H., Banks, G. C. and Gonzalez-Mulé, E.: 2017, The chrysalis effect: How ugly initial results metamorphosize into beautiful articles, *Journal of Management* **43**(2), 376–399.

Open Science Collaboration: 2015, Estimation the reproducibility of psychological science, *Science* **349**(6251), aac1716.

Pinheiro, J. C. and Bates, D. M.: 2000, *Mixed Effects Models in S and S-PLUS*, Springer.

Qian, J. and Wang, L.: 2012, Estimating semiparametric panel data models by marginal integration, *Journal of Econometrics* **167**(2), 483–493.

Rigby, R. A. and Stasinopoulos, D. M.: 2005, Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3), 507–554.

Rojas-Perilla, N., Pannier, S., Schmid, T. and Tzavidis, N.: 2017, Data-driven transformations in small area estimation. Discussion Paper 30/2017, School of Business and Economics, Freie Universität Berlin, available from `https://www.econstor.eu/bitstream/10419/172326/1/1009227645.pdf`.

Rosenthal, R.: 1979, The file drawer problem and tolerance for null results, *Psychological Bulletin* **86**(3), 638–641.

Rothe, C.: 2010, Nonparametric estimation of distributional policy effects, *Journal of Econometrics* **155**(1), 56–70.

Rothe, C.: 2012, Partial Distributional Policy Effects, *Econometrica* **80**(5), 2269–2301.

Ruppert, D. and Wand, M.: 2003, *Semiparametric Regression*, Cambridge University Press.

Schnabel, S. K. and Eilers, P. H. C.: 2013, Simultaneous estimation of quantile curves using quantile sheets, *AStA Advances in Statistical Analysis* **97**(1), 77–87.

Shen, S.: 2019, Estimation and inference of distributional partial effects: Theory and application, *Journal of Business & Economic Statistics* **37**(1), 54–66.

Shen, S. and Zhang, X.: 2016, Distributional tests for regression discontinuity: Theory and empirical examples, *The Review of Economics and Statistics* **98**(4), 685–700.

Simmons, J. P., Nelson, L. D. and Simonsohn, U.: 2011, False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant, *Psychological Science* **22**(11), 1359–1366.

Simonsohn, U.: 2015, Small telescopes: Detectability and the evaluation of replication results, *Psychological science* **26**(5), 559–569.

Simonsohn, U., Nelson, L. D. and Simmons, J. P.: 2014, P-curve: A key to the file-drawer, *Journal of Experimental Psychology: General* **143**(2), 534–547.

Sobotka, F., Radice, R., Marra, G. and Kneib, T.: 2013, Estimating the relationship between women's education and fertility in Botswana by using an instrumental variable approach to semiparametric expectile regression, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**(1), 25–45.

Stasinopoulos, M. D. and Rigby, R. A.: 2007, Generalized additive models for location scale and shape (GAMLSS) in R, *Journal of Statistical Software* **23**(7).

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. and De Bastiani, F.: 2017, *Flexible Regression and Smoothing: Using GAMLSS in R*, CRC Press.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. and Robinson, G. E.: 2015, Big data: Astronomical or genomical?, *PLoS biology* **13**(7), e1002195.

Stock, J. H. and Watson, M. W.: 2011, *Introduction to econometrics (third edition)*, Addison-Wesley.

Su, L. and Ullah, A.: 2006, Profile likelihood estimation of partially linear panel data models with fixed effects, *Economics Letters* **92**, 75–81.

Su, L. and Ullah, A.: 2011, Nonparametric and semiparametric panel econometric models: Estimation and testing, in *Handbook of Empirical Economics and Finance*, Taylor & Francis Group, pp. 455–497.

Sun, J. and Loader, C. R.: 1994, Simultaneous confidence bands for linear regression and smoothing, *The Annals of Statistics* **22**(3), 1328–1345.

Tarozzi, A. and Deaton, A.: 2009, Using census and survey data to estimate poverty and inequality for small areas, *Review of Economics and Statistics* **91**(4), 773–792.

Terza, J. V., Basu, A. and Rathouz, P. J.: 2008, Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling, *Journal of Health Economics* **27**(3), 531–543.

The National Statistical Coordination Board of the Philippines: 2009, 2003 city and municipal level

poverty estimates. Available from https://psa.gov.ph/sites/default/files/2003%20SAE%20of%20poverty%20%28Full%20Report%29_0.pdf.

The Bangladesh Bureau of Statistics, The World Bank and The United Nations World Food Programme: 2010, Updating poverty maps: Bangladesh poverty maps for 2005. Available from http://www.wfp.org/sites/default/files/Poverty%20Map%202005%20Technical%20Report.pdf.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K. and Schinke, S. P.: 2011, Replication in prevention science, *Prevention Science* **12**(2), 103–117.

Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., Van Assen, M. A. and Wicherts, J. M.: 2014, Statistical reporting errors and collaboration on statistical analyses in psychological science, *PLoS ONE* **9**(12), 1–19.

Vivalt, E.: 2019, Specification searching and significance inflation across time, methods and disciplines, *Oxford Bulletin of Economics and Statistics* (forthcoming).

Wager, S. and Athey, S.: 2018, Estimation and inference of heterogeneous treatment effects using random forests, *Journal of the American Statistical Association* **113**(523), 1228–1242.

Wagner, G. G., Frick, J. R. and Schupp, J.: 2007, The German socio-economic panel study (SOEP) - scope, evolution, and enhancements, *Schmollers Jahrbuch* **127**(1), 139–169.

Wang, J. and Yang, L.: 2009, Polynomial spline confidence bands for regression curves, *Statistica Sinica* **19**, 325–342.

Wang, X., Shen, J. and Ruppert, D.: 2011, On the asymptotics of penalized spline smoothing, *Electronic Journal of Statistics* **5**, 1–17.

Wasserstein, R. L., Lazar, N. A. et al.: 2016, The asa's statement on p-values: context, process, and purpose, *The American Statistician* **70**(2), 129–133.

Weyl, H.: 1939, On the volume of tubes, *American Journal of Mathematics* **61**(2), 461–472.

Wicherts, J. M., Bakker, M. and Molenaar, D.: 2011, Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results, *PLoS ONE* **6**(11), 1–7.

Wiesenfarth, M., Krivobokova, T., Klasen, S. and Sperlich, S.: 2012, Direct simultaneous inference in additive models and its application to model undernutrition, *Journal of the American Statistical Association* **107**(500), 1286–1296.

Wood, S. N.: 2006, *Generalized Additive Models: An introduction with R*, Chapman & Hall/CRC.

Wood, S. N., Pya, N. and Säfken, B.: 2016, Smoothing parameter and model selection for general smooth models, *Journal of the American Statistical Association* **1459**(8), 1–45.

Wooldridge, J. M.: 2002, *Econometric Analysis of Cross Section and Panel Data (second edition)*, MIT Press.

Wooldridge, J. M.: 2014, Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory variables, *Journal of Econometrics* **182**(1), 226–234.

Wunder, C., Wiencierz, A., Schwarze, J. and Küchenhoff, H.: 2011, Well-being over the life span: Semi-parametric evidence from British and German longitudinal data, *Review of Economics and Statistics* **95**(1), 154–167.

Yang, L.: 2008, Confidence band for additive regression model, *Journal of Data Science* **6**(2), 207–217.

Yoshida, T. and Naito, K.: 2012, Asymptotics for penalized additive B-spline regression, *Journal of the Japan Statistical Society* **42**(1), 81–107.

Yoshida, T. and Naito, K.: 2014, Asymptotics for penalized splines in generalized additive models, *Journal of Nonparametric Statistics* **26**(2), 269–289.

Zhang, J., Feng, S., Li, G. and Lian, H.: 2011, Empirical likelihood inference for partially linear panel data models with fixed effects, *Economics Letters* **113**(2), 165–167.

## Promotionsstudiengang „Wirtschaftswissenschaften“
## Versicherung bei Zulassung zur Promotionsprüfung

Ich versichere,

1. dass ich die eingereichte Dissertation
   *Causality, Prediction, and Replicability in Applied Statistics: Advanced Models and Practices*
   selbstständig angefertigt habe und nicht die Hilfe Dritter in einer dem Prüfungsrecht und wissenschaftlicher Redlichkeit widersprechenden Weise in Anspruch genommen habe,
2. dass ich das Prüfungsrecht einschließlich der wissenschaftlichen Redlichkeit – hierzu gehört die strikte Beachtung des Zitiergebots, so dass die Übernahme fremden Gedankenguts in der Dissertation deutlich gekennzeichnet ist – beachtet habe,
3. dass beim vorliegenden Promotionsverfahren kein Vermittler gegen Entgelt eingeschaltet worden ist sowie im Zusammenhang mit dem Promotionsverfahren und seiner Vorbereitung
   - kein Entgelt gezahlt oder entgeltgleiche Leistungen erbracht worden sind
   - keine Dienste unentgeltlich in Anspruch genommen wurden, die dem Sinn und

   Zweck eines Prüfungsverfahrens widersprechen
4. dass ich eine entsprechende Promotion nicht anderweitig beantragt und hierbei die eingereichte Dissertation oder Teile daraus vorgelegt habe.

Mir ist bekannt, dass Unwahrheiten hinsichtlich der vorstehenden Versicherung die Zulassung zur Promotionsprüfung ausschließen und im Falle eines späteren Bekanntwerdens die Promotionsprüfung für ungültig erklärt werden oder der Doktorgrad aberkannt werden kann.

Datum, Unterschrift