Treatment effects beyond the mean using GAMLSS

Maike Hohberg [*1], Peter Pütz[1], and Thomas Kneib[1]

[1]*University of Goettingen*

March 4, 2019

Summary

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 nonnormal 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

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), analysing 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 2.

Other interesting approaches to go beyond the mean in regression modelling 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 (2017) proposed a non-parametric 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 (e.g. 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 2 provides the methodological background of

GAMLSS. Section 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 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 5 concludes.

2 Generalized additive models for location, scale and shape

2.1 A general introduction to GAMLSS

For the sake of illustration, we start with a basic regression as it would be for example the case 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 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

$$y_i = \beta_0 + \beta_T T_i + \mathbf{x}'_i \boldsymbol{\beta}_1 + \varepsilon_i \tag{1}$$

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

$$E(y_i) = \mu_i = \beta_0 + \beta_T T_i + \mathbf{x}'_i \boldsymbol{\beta}_1. \tag{2}$$

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, i.e. 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:

$$\mu_i = \beta_0^\mu + \beta_T^\mu T_i + \mathbf{x}'_i \boldsymbol{\beta}_1^\mu, \tag{3}$$

$$\log(\sigma_i) = \beta_0^\sigma + \beta_T^\sigma T_i + \mathbf{x}'_i \boldsymbol{\beta}_1^\sigma. \tag{4}$$

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 (4) 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, i.e. where the response is a vector of dependent variables, can be placed within this modeling framework (Klein et al., 2015a).

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

$$g_k(\vartheta_{ik}) = \beta_0^{\vartheta_k} + \beta_T^{\vartheta_k} T_i + x_i' \boldsymbol{\beta}^{\vartheta_k}, \tag{5}$$

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)). Linking the parameters to an unconstrained domain also facilitates the consideration of semiparametric, additive regression specifications incuding 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 et al., 2015c).

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 ϑ_{ik} takes on the following form:

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + \beta_T^{\vartheta_k} T_i + f_1^{\vartheta_k}(\mathbf{x}_{1i}) + \cdots + f_{J_k}^{\vartheta_k}(\mathbf{x}_{J_k i}). \tag{6}$$

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 is 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).

2.3 GAMLSS vs. quantile regression

A popular alternative to simple mean regression is quantile regression, see, e.g., 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

$$y_i = \beta_{0,\tau} + \beta_{1,\tau} x_i + \varepsilon_{i,\tau} \tag{7}$$

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 τ-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, e.g. based on a location scale shift model (He, 1997), on spline based noncrossing 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 separatetely 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 et al. (2015b) show bad performance for likelihood-based confidence intervals in certain situations, we will, however, rely on bootstrap inference for the application in Section 4. The main drawback of GAMLSS is a potential misspecification but Section 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 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. 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 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 non-negative 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, i.e. 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 et al. (2015a) for an application.

Finally, as shown in Section 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 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 is 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 (e.g. continuous, non-negative, binary, discrete) 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

$$\ln y_i = \beta_0^\mu + \mathbf{x}_i'\boldsymbol{\beta}_1^\mu + \varepsilon_i, \tag{8}$$

where y_i is consumption or income, β_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 ε_i is a normally distributed error term with variance

$$\sigma_{e,i}^2 = \beta_0^\sigma + \mathbf{x}_i'\boldsymbol{\beta}_1^\sigma. \tag{9}$$

To estimate the intercepts β_0^μ and β_0^σ 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

$$\widehat{\Pr}(\ln y_i < \ln z | \mathbf{x}_i') = \Phi\left(\frac{\ln z - (\hat{\beta}_0^\mu + \mathbf{x}_i'\hat{\boldsymbol{\beta}}_1^\mu)}{\sqrt{\hat{\beta}_0^\sigma + \mathbf{x}_i'\hat{\boldsymbol{\beta}}_1^\sigma}}\right), \tag{10}$$

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 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, i.e. 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, e.g. 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. The fair notion would be that a person with higher education should earn more. It is the conditional inequality that is perceived as unfair.

4   Applying GAMLSS to experimental data

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, e.g., 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.

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 (nonpoor) 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 nonpoor, 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, e.g. whether conditional food consumption inequality decreases for an average household among the poor (or the nonpoor or all households). We will assess the effect on conditional inequality via the Gini coefficient, which is in general defined by

$$
G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|}{2n \sum_{h=1}^{n} y_i}, \ 0 \leq G \leq 1,
\tag{11}
$$

for a group of n households, where y_i denotes the non-negative 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

$$
G = \frac{1}{2\mu} \int_0^\infty \int_0^\infty p(y)p(z) \, |y - z| \, dy \, dz,
\tag{12}
$$

with μ 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 (11) and remain the same, but the denominator increases. An equivalent logic applies to (12). However, there as well might be 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 nonpoor 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 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, i.e. 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 nonpoor, a lower food consumption inequality (measured by the Gini coefficent) 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 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 10000 pesos per adult equivalent. Along the steps described in Section 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`.

4.2.1 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, see Section 4.2.3.

The histogram of the dependent variable in the left panel of Figure 1 shows a heavily right-skewed distribution.

Figure 1: Histogram of food consumption and log food consumption
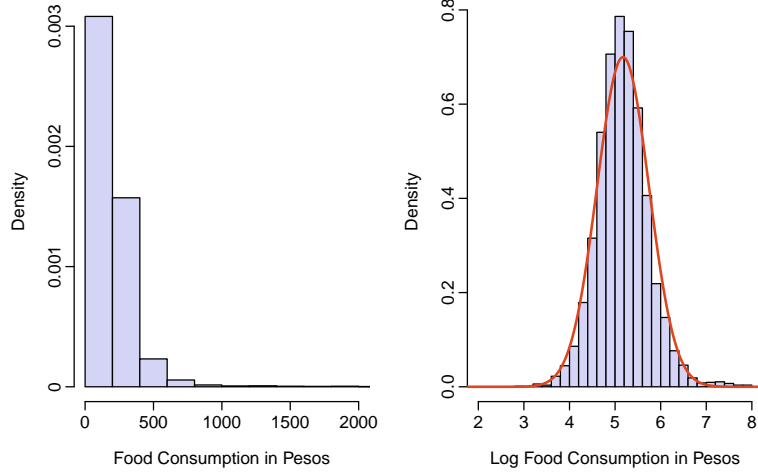
The logarithm of the dependent variable in the right panel of Figure 1 somewhat resembles a normal distribution such that the log-normal distribution appears to be a reasonable starting point. It has the additional advantage that it also renders easily interpretable effects of the explanatory variables 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.

4.2.2 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 μ and σ are related to these variables,

$$\log(\mu_i) = \beta_0^\mu + T_i\beta_T^\mu + \mathbf{x}_i'\boldsymbol{\beta}_1^\mu, \tag{13}$$

$$\log(\sigma_i) = \beta_0^\sigma + T_i\beta_T^\sigma + \mathbf{x}_i'\boldsymbol{\beta}_1^\sigma, \tag{14}$$

where T_i is the treatment dummy, β_T^μ and β_T^σ are the treatment effects on the parameters μ and σ, 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 μ and σ are modeled as in (13) and (14), respectively, an additional parameter τ is linked to the nine explanatory variables,

$$\log(\tau_i) = \beta_0^\tau + T_i\beta_T^\tau + \mathbf{x}_i'\boldsymbol{\beta}_1^\tau, \tag{15}$$

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, i.e. 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.

4.2.3 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 2a and Table 1 it can be seen that both qq-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.
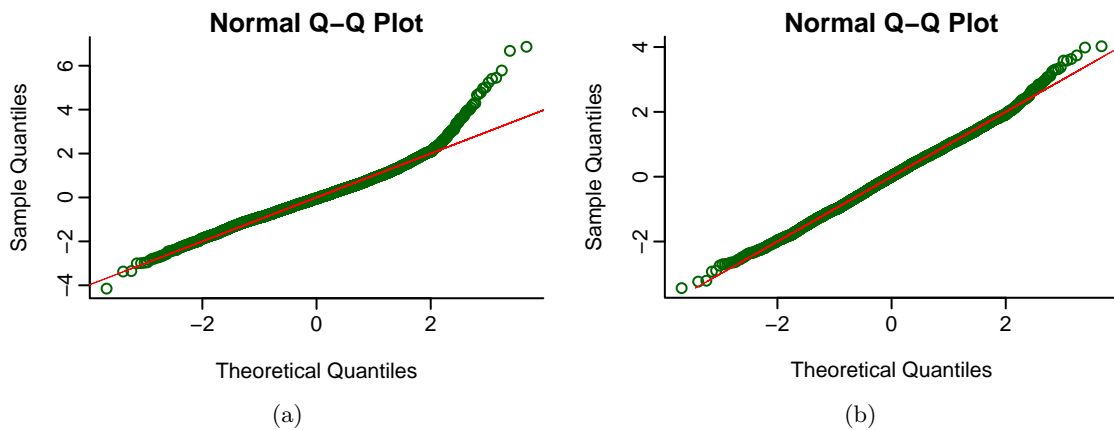

Figure 2: Diagnosis plots for the model based on (a) the log-normal distribution and (b) the Singh-Maddala distribution

14

Table 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

Note: 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.

In contrast, a model relying on the Singh-Maddala distribution yields a much more satisfying diagnostic fit (see Figure 2b and Table 1). The qq-plot does not show severe deviations from the standard normal distribution, which is confirmed by the summary measures of the quantile residuals. More specifically, 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, e.g. 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).

4.2.4   Variable selection

A comparison between different models, for instance our model of choice from above to 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.

4.2.5 Reporting and interpreting the results

GAMLSS using the Singh-Maddala distribution relate three parameters (via link functions) non-linearly 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, i.e. 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.

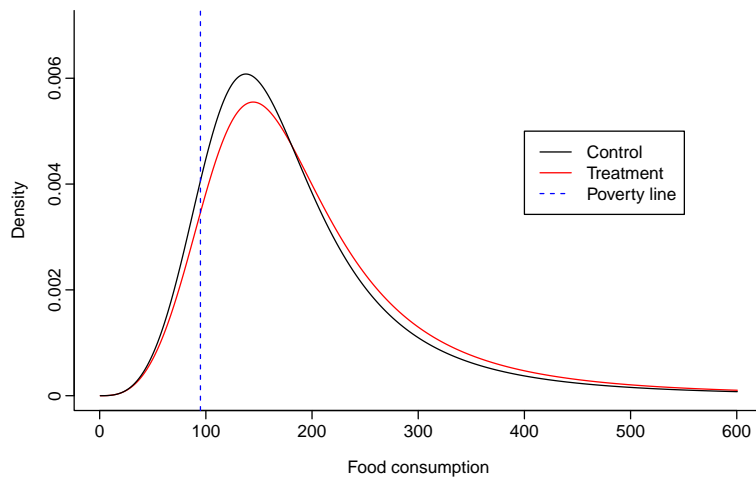Figure 3: Estimated conditional distributions for an average household

Table 2: Treatment effects for ineligibles ($n = 4248$): Point estimates and 95% bootstrap confidence interval bounds for marginal treatment effects at means (MTE)

	Estimate	Lower Bound	Upper Bound
MTE on mean	16.232	2.343	28.556
MTE on variance	8463.007	-5842.096	22787.013
MTE on Gini coefficient	0.014	-0.011	0.036
MTE on Atkinson index (e=1)	0.012	-0.009	0.031
MTE on Atkinson index (e=2)	0.018	-0.011	0.046
MTE on Theil index	0.019	-0.020	0.051
MTE on vulnerability	-0.015	-0.038	0.008

The results in Table 2 show point estimates and 95% bootstrap percentile intervals of marginal treatment effects for an average household, i.e. 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 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 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: Treatment effects for eligibles ($n = 10492$): Point estimates and 95% bootstrap confidence interval bounds for marginal treatment effects at means (MTE)

	Point Estimate	Lower Bound	Upper Bound
MTE on mean	28.898	19.970	37.781
MTE on variance	4561.613	406.903	8696.840
MTE on Gini	0.007	-0.010	0.021
MTE on Atkinson (e=1)	0.006	-0.008	0.017
MTE on Atkinson (e=2)	0.012	-0.009	0.028
MTE on Theil-Index	0.007	-0.015	0.026
MTE on Vulnerability	-0.077	-0.107	-0.051

Qualitatively the same results emerge for the group of eligibles, as can be seen in Table 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 4,

Table 4: Treatment effects for all people in treatment villages ($n = 14740$): Point estimates and 95% bootstrap confidence interval bounds for marginal treatment effects at means (MTE)

	Point Estimate	Lower Bound	Upper Bound
MTE on mean	25.907	19.571	32.251
MTE on variance	4870.407	1703.281	8852.551
MTE on Gini	0.007	-0.004	0.021
MTE on Atkinson (e=1)	0.006	-0.003	0.017
MTE on Atkinson (e=2)	0.012	-0.002	0.028
MTE on Theil-Index	0.007	-0.008	0.025
MTE on Vulnerability	-0.056	-0.073	-0.039

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, 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.

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 advantages of GAMLSS are 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 nonpoor that were larger for the poor (the target group) than for the nonpoor. 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 data sets 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.

References

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.

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.

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 Bayesian Inference in Structured Additive Regression Models. Version 3.0.2. Available from `http://www.bayesx.org`.

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.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2016). Regression Discontinuity Designs Using Covariates (Working paper, University of Michigan). Available from `https://www.arec.umd.edu/sites/arec.umd.edu/files/_docs/events/Matias%20Cattaneo-Regression%20Discontinuity%20Designs%20Using%20Covariates.pdf`.

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*, 50:317–372.

Celidoni, M. (2013). Vulnerability to Poverty: An Empirical Comparison of Alternative Measures. *Applied Economics*, 45:1493–1506.

Chaudhuri, S., Jalan, J., and Suryahadi, A. (2002). Assessing Household Vulnerability to Poverty from Cross-sectional Data: A Methodology and Estimates from Indonesia. Technical Report 0102-52, Department of Economics, Columbia University, New York.

Chavez-Martin del Campo, J. (2006). Does Conditionality Generate Heterogeneity and Regressivity in Program Impacts? The Progresa Experience. Technical report, Cornell University, Department of Applied Economics and Management.

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*, (1803.08154).

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.

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.

Dunn, P. K. and Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5(3):236.

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):89–121.

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

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.

Frandsen, B. R., Frölich, M., and Melly, B. (2012). Quantile Treatment Effects in the Regression Discontinuity Design. *Journal of Econometrics*, 168(2):382–395.

Frölich, M. and Melly, B. (2013). Unconditional Quantile Treatment Effects Under Endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357.

Geraci, A., Fabbri, D., and Monfardini, C. (2016). Testing Exogeneity of Multinomial Regressors in Count Data Models: Does Two-stage Residual Inclusion Work? *Journal of Econometric Methods*, 7(1).

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

Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, 46(4):931.

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).

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.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences.* Wiley, Hoboken.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2015a). 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. (2015b). 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. (2015c). 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, Cambridge.

Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1):33.

Lee, D. S. and Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48:281–355.

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.

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.

Meager, R. (2016). Aggregating Distributional Treatment Effects: A Bayesian Hierarchical Analysis of the Microcredit Literature. 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.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica: journal of the Econometric Society*, 46(1):69–85.

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.

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.

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. (2017). Estimation and Inference of Distributional Partial Effects: Theory and Application. *Journal of Business & Economic Statistics*, pages 1–13.

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.

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.

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.

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.* MIT Press, Boston, Ma, 2nd edition.

Wooldridge, J. M. (2014). Quasi-Maximum Likelihood Estimation and Testing for Nonlinear Models with Endogenous Explanatory variables. *Journal of Econometrics*, 182(1):226–234.

Appendix

A Combining evaluation methods for non-experimental data and GAMLSS

As demonstrated in Section 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.

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:

$$y_{it} = \beta_0 + \mathbf{x}'_{it}\boldsymbol{\beta}_1 + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T_i. \tag{16}$$

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 α_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 a 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

$$y_{it} = \beta_0 + \mathbf{x}'_{it}\boldsymbol{\beta}_1 + \bar{\mathbf{x}}'_i\boldsymbol{\delta}_1 + \alpha_i + \varepsilon_{it}, \quad i = 1, \dots, N,\ t = 1, \dots, T_i, \tag{17}$$

where α_i, $i = 1, \dots, N$, 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 (17) is equivalent to $\boldsymbol{\beta}_1$ in a fixed effects model according to (16).

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, i.e 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 (17), we can specify for each of these parameters an equation of the form

$$g_k(\vartheta_{itk}) = \beta_0^{\vartheta_k} + \mathbf{x}'_{it}\boldsymbol{\beta}_1^{\vartheta_k} + \bar{\mathbf{x}}'_i\boldsymbol{\delta}_1^{\vartheta_k} + \alpha_i^{\vartheta_k}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T_i, \tag{18}$$

with link function g_k, see Sections 2.1 and 2.2 in the main text for details and extensions.

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

$$y = \beta_0 + x_e\beta_e + x_o\beta_o + x_u\beta_u + \varepsilon \quad \text{with} \quad E(\varepsilon|x_e, x_o, x_u) = 0, \tag{19}$$

where x_o is an observed explanatory variable, x_e the endogenous variable, x_u the unobserved confounder, ε is an error term and β_o, β_e, and β_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 β_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.

A.2.1 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

$$E(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u) = h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \mathbf{X}_u\boldsymbol{\beta}_u), \tag{20}$$

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 (20) can be written as

$$\mathbf{y} = h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \mathbf{X}_u\boldsymbol{\beta}_u) + \boldsymbol{\varepsilon} \tag{21}$$

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

$$E(\boldsymbol{\varepsilon}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u) = \mathbf{0}. \tag{22}$$

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 (21) 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

$$\mathbf{x}_{es} = h_s(\mathbf{X}_o\boldsymbol{\alpha}_{os} + \mathbf{W}_s\boldsymbol{\alpha}_{ws}) + \boldsymbol{\xi}_{us}, \qquad s = 1, \ldots, S_e \tag{23}$$

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 (23) 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, i.e. $\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u$

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

Terza et al. (2008) propose the following procedure to estimate the models in equations (21) and (23):

(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 (23) via a consistent estimation strategy. One could use maximum likelihood estimation for GLMs here, but nonlinear least squares is also possible. Define

$$\hat{\boldsymbol{\xi}}_{us} = \mathbf{x}_{es} - h(\mathbf{X}_o\hat{\boldsymbol{\alpha}}_{os} + \mathbf{W}_s\hat{\boldsymbol{\alpha}}_{ws}) \qquad \text{for} \quad s = 1, \ldots, S_e. \tag{24}$$

(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

$$E(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \hat{\boldsymbol{\Xi}}_u) = h(\mathbf{X}_e\boldsymbol{\beta}_e + \mathbf{X}_o\boldsymbol{\beta}_o + \hat{\boldsymbol{\Xi}}_u\boldsymbol{\beta}_{\hat{\Xi}_u}), \tag{25}$$

25

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).

A.2.2 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

$$\mathbf{y} = h(\boldsymbol{\eta}) + \boldsymbol{\varepsilon} \qquad E(\boldsymbol{\varepsilon}|\mathbf{X}_e, \mathbf{X}_o, \mathbf{X}_u) = 0, \tag{26}$$

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 $\mathbf{X}^+$, i.e. $\mathbf{X}^* = (\mathbf{X}_e^*, \mathbf{X}_o^*, \mathbf{X}_u^*)$ 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

$$\boldsymbol{\eta} = \mathbf{X}^*\boldsymbol{\beta}^* + \sum_{l=1}^{L} f_l(\mathbf{x}_l^+), \tag{27}$$

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

$$\mathbf{x}_{es} = h_s(\mathbf{Z}_s^*\boldsymbol{\alpha}_s^* + \sum_{j=1}^{J_s} f_j(\mathbf{z}_{js}^+)) + \boldsymbol{\xi}_{us}, \tag{28}$$

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 (27) and f_j (28) 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 (28) using a GAM method. Define

$$\hat{\boldsymbol{\xi}}_{us} = \mathbf{x}_{es} - h_s(\mathbf{Z}_s^*\hat{\boldsymbol{\alpha}}_s^* + \sum_{j=1}^{J_s} \hat{f}_j(\mathbf{z}_{js}^+)) \qquad \text{for} \quad s = 1, \ldots, S_e. \tag{29}$$

(b) Second stage: Estimate

$$E(\mathbf{y}|\mathbf{X}_e, \mathbf{X}_o, \hat{\boldsymbol{\Xi}}_u) = h_s(\mathbf{X}_e^* \boldsymbol{\beta}_e^* + \mathbf{X}_o^* \boldsymbol{\beta}_o^* + \sum_{j=1}^{J} f_j(\mathbf{x}_{jeo}^+) + \sum_{s=1}^{S_e} f_s(\hat{\boldsymbol{\xi}}_{us})), \qquad (30)$$

where $\mathbf{x}_{jeo}^+, j = 1, \dots, J$, 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 (29) and (30) 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.

A.2.3 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 $\boldsymbol{\vartheta} = (\vartheta_1, \dots, \vartheta_K)'$ and density

$$p(\mathbf{y}|\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u) = p(\mathbf{y}|\boldsymbol{\vartheta}(\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u)) \qquad (31)$$

For each of the parameters, a regression specification

$$\vartheta_k = h_k(\eta^{\vartheta_k}) \qquad (32)$$

is assumed, where η^{ϑ_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

$$\boldsymbol{\eta}^{\vartheta_k}(\mathbf{X}_o, \mathbf{X}_e, \mathbf{X}_u) = \mathbf{X}^* \boldsymbol{\beta}^{*,\vartheta_k} + \sum_{l=1}^{L} f_l^{\vartheta_k}(\mathbf{x}_l^+) \qquad (33)$$

Using the same notation as above, the only difference between the equations (27) and (33) 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 (33) 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

$$\boldsymbol{\eta}^{\vartheta_k} = \mathbf{X}_e^* \boldsymbol{\beta}_e^{*,\vartheta_k} + \mathbf{X}_o^* \boldsymbol{\beta}_o^{*,\vartheta_k} + \sum_{j=1}^{J} f_j^{\vartheta_k}(\mathbf{x}_{jeo}^+) + \sum_{s=1}^{S_e} f_s^{\vartheta_k}(\hat{\boldsymbol{\xi}}_{us}). \qquad (34)$$

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 $E(\mathbf{x}_{es}|\mathbf{X}_o, \mathbf{W}_s), s = 1, \dots, S_e$, 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

$$p(\mathbf{X}_u | \mathbf{X}_o, \mathbf{x}_{es}, \mathbf{W}_s) = p(\mathbf{X}_u | C_s(\mathbf{X}_o, \mathbf{x}_{es}, \mathbf{W}_s)). \tag{35}$$

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

$$\mathbf{x}_{es} | \mathbf{X}_o, \mathbf{W}_s \sim N(\boldsymbol{\eta}^{\vartheta_k}(\mathbf{X}_o, \mathbf{W}_s), \boldsymbol{\sigma}_{es}^2), \tag{36}$$

then

$$\boldsymbol{\xi}_u = \mathbf{x}_{es} - \mathbf{Z}_s^* \boldsymbol{\alpha}_s^* + \sum_{j=1}^{J_s} f_j(\mathbf{z}_{js}^+) \tag{37}$$

is an appropriate control function in the sense that assumption (35) 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 e.g. 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 (35) 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 $\hat{\boldsymbol{\xi}}_u$ nonlinearly and/or interactions with $\mathbf{X}_e, \mathbf{X}_o$ in (34) to improve the approximation. Furthermore, a simulation study on different 2SRI settings suggested standardizing the variance of the first stage residuals (Geraci et al., 2016).

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 B.3.

A.3   Regression discontinuity design

In the regression discontinuity design (RDD), e.g. see 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

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \text{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \text{E}[Y_i | X_i = x], \tag{38}$$

where Y_i is the dependent variable of interest. The two quantities in (38) 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 $E[Y_i|X_i, X_i > c]$ and $E[Y_i|X_i, X_i < c]$ 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 for instance binary. Hereby, the crucial assumption is the continuity in the counterfactual conditional mean functions $E[Y_i(0)|X_i = x]$ and $E[Y_i(1)|X_i = x]$, 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 (38) 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 $\vartheta_{i1}, \ldots, \vartheta_{iK}$, where K are 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

$$g_k(\vartheta_{ik}) = \beta_0^{\vartheta_k} + X_i \beta_2^{\vartheta_k}, \quad i = 1, \ldots, N, \tag{39}$$

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. (2016) 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 seemingly 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, e.g. 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, i.e. 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

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \ \text{E}[Y|X_i = x] - \lim_{x \uparrow c} \ \text{E}[Y|X_i = x]}{\lim_{x \downarrow c} \Pr(T_i = 1|X_i = x) - \lim_{x \uparrow c} \Pr(T_i = 1|X_i = x)}, \tag{40}$$

where the denominator now includes the probabilities of treatment at both sides near the cutoff. The treatment effect in (40) 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 (40) 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, i.e. that individuals do not become ineligible for the treatment or discouragend 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 (40) may also contain differences in other quantities aside from the conditional means. The probabilities in the denominator in (40) can be estimated separately, e.g. 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.

B   Bootstrap inference

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

B.1   General strategy

To fix ideas, assume without loss of generality that the quantity of interest is the marginal treatment effect at the means, i.e. for an averag e individual, on the Gini coefficient, denoted by θ. 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 data set 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 nontreated 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

$$\hat{V}_{\text{boot}}[\hat{\theta}] = \frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\hat{\theta}})^2 \tag{41}$$

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

$$t = \frac{\hat{\theta}}{\sqrt{\hat{V}_{\text{boot}}[\hat{\theta}]}} \tag{42}$$

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 2.5 and the upper 97.5 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, for instance Efron (1987) 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.

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 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 nontreated 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 nonclustered data, a bootstrap t-test can be conducted with the denominator in (42) now based on the cluster-robust variance estimator

$$\hat{V}_{\text{clu;boot}}[\hat{\theta}] = \frac{c}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\hat{\theta}})^2, \tag{43}$$

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 B.1.

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 B.1 for the first stage model in equation 28.

(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

$$\hat{\mathbf{x}}_{es}^{[k]} = h(\mathbf{Z}_s \hat{\boldsymbol{\alpha}}_s^{[k]}) \tag{44}$$

and

$$\hat{\boldsymbol{\xi}}_{us}^{[k]} = \mathbf{x}_{es} - \hat{\mathbf{x}}_{es}^{[k]}. \tag{45}$$

(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 B.1.

As an alternative to the parametric bootstrap in step 1, a non-parametric 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.

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 (40). By doing so, the uncertainty of those estimates is included in the resulting standard errors or confidence intervals for the treatment effect of interest.

B.5 Recommendations for diagnosing bootstrap estimates

Irrespective of the impact evaluation and bootstrap method chosen, but especially in the case of the pairs 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 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 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 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.

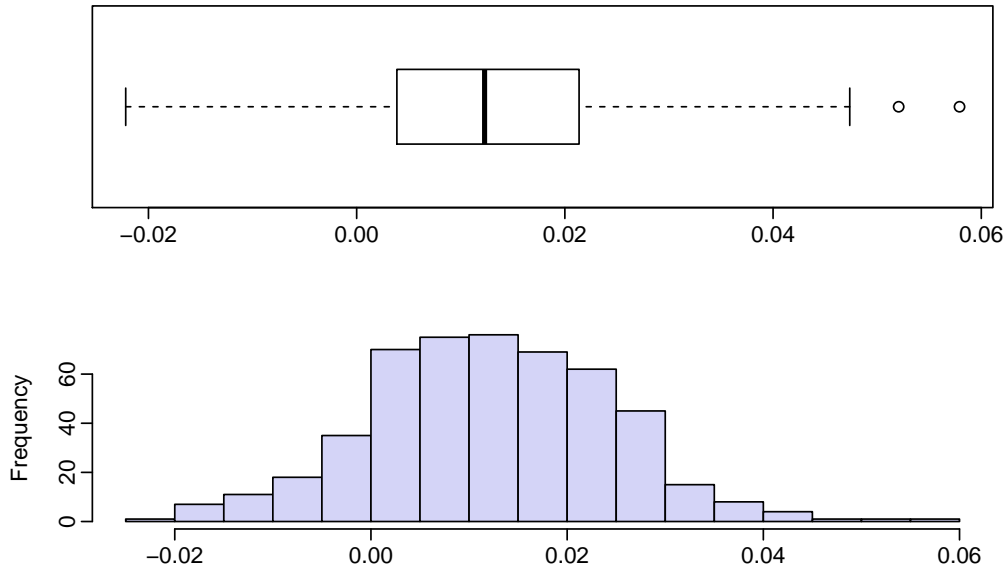Figure 4: Distribution of bootstrap estimates of MTE on Gini

Figure 5: Percentile Interval bounds for MTE on Gini for increasing bootstrap replicates