Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 4, *pp.* 1553–1574

Flexible instrumental variable distributional regression

Guillermo Briseño Sanchez,

TU Dortmund University, Germany

Maike Hohberg,

University of Göttingen, Germany

Andreas Groll

TU Dortmund University, Germany

and Thomas Kneib

University of Göttingen, Germany

[Received October 2018. Final revision June 2020]

Summary. We tackle two limitations of standard instrumental variable regression in experimental and observational studies: restricted estimation to the conditional mean of the outcome and the assumption of a linear relationship between regressors and outcome. More flexible regression approaches that solve these limitations have already been developed but have not yet been adopted in causality analysis. The paper develops an instrumental variable estimation procedure building on the framework of generalized additive models for location, scale and shape. This enables modelling all distributional parameters of potentially complex response distributions and non-linear relationships between the explanatory variables, instrument and outcome. The approach shows good performance in simulations and is applied to a study that estimates the effect of rural electrification on the employment of females and males in the South African province of KwaZulu-Natal. We find positive marginal effects for the mean for employment of females rates, negative effects for employment of males and a reduced conditional standard deviation for both, indicating homogenization in employment rates due to the electrification programme. Although none of the effects are statistically significant, the application demonstrates the potentials of using generalized additive models for location, scale and shape in instrumental variable regression for both to account for endogeneity and to estimate treatment effects beyond the mean.

Keywords: Causality; Distributional regression; Generalized additive models for location, scale and shape; Instrumental variable; Treatment effects

1. Introduction

In the potential outcomes framework (Neyman, 1990; Rubin, 1974), causal effects of a binary treatment $D \in \{0, 1\}$ on an outcome variable of interest $Y(D)$ are defined as comparisons between the potential outcome under treatment, $Y_1 = Y(1)$, and the potential outcome without treatment, $Y_0 = Y(0)$, for a common set of units. In practice, we shall be able to observe only either Y_1 or Y_0, depending on the treatment status. When the effect of the treatment is restricted to the mean

Address for correspondence: Guillermo Briseño Sanchez, Fakultät Statistik, Technische Universität Dortmund, Vogelspothsveg 87, Dortmund, Nordhein Westfalen 44227, Germany.
E-mail: briseno@statistik.tu-dortmund.de

of the outcome, the causal effect of the treatment can be reduced to one scalar quantity: the average treatment effect (ATE)

$$\text{ATE} = E(Y_1) - E(Y_0).$$

In heterogeneous populations, the ATE is usually extended to depend on a characteristic X of the individuals of interest, leading to the conditional ATE

$$\text{ATE}(x) = E(Y_1|X = x) - E(Y_0|X = x).$$

Later, we shall consider several characteristics but, for this introductory section, we restrict the notation to the univariate case. The problem with the scalar ATE is that it provides only a rather narrow view of the treatment effect.

Policy makers are often concerned with questions that relate to more general distributional aspects of the variable of interest, such as income inequality, and might then prefer an intervention that lowers the variance or the Gini coefficient of an income distribution over an intervention that has the same ATE but does not reduce inequality. Thus, a more general perspective is to consider potential changes in the complete (conditional) distribution of the outcome, i.e. differences between $\mathcal{D}(Y_1|X = x)$ and $\mathcal{D}(Y_0|X = x)$. In this case, there is not one single scalar treatment effect but rather several treatment effects on various aspects of the distribution of the outcome. This has received considerable interest in the context of quantile regression where conditional treatment effects on specific quantiles $\tau \in (0, 1)$ of the distribution of the outcome can be defined as quantile treatment effects with

$$\text{QTE}_\tau(x) = Q_{Y_1|X=x}(\tau) - Q_{Y_0|X=x}(\tau), \tag{1}$$

where $Q_{Y|X=x|}(\cdot)$ refers to the quantile function of the distribution of the treatment and control potential outcome (see Melly and Wüthrich (2017) for a review of quantile treatment effects in the context of IVs). However, if interest is not only on a specific quantile but also on different features of the distribution such as the variance or the Gini coefficient, it is desirable to estimate the whole conditional distribution directly. With quantile regression, this would require estimating numerous quantile effects and often also dealing with the problem of crossing quantiles.

In this paper, we consider a different approach to evaluate the difference between $\mathcal{D}(Y_1|X = x)$ and $\mathcal{D}(Y_0|X = x)$, where a parametric type of distribution is assumed for Y such that, for example, $\mathcal{D}(Y_1|X = x) = \mathcal{D}_{\vartheta_{Y_1}(x)}$ and $\mathcal{D}(Y_0|X = x) = \mathcal{D}_{\vartheta_{Y_0}(x)}$. This assumes the same type of distribution for Y with and without treatment whereas the parameters of the distribution differ by the treatment (and covariates). One can then either evaluate differences in the parameters with and without treatment directly or derive certain other quantities of interest from the parameters.

To illustrate this point in more detail, assume that $Y_1|X = x \sim \mathcal{N}\{\mu_{Y_1}(x), \sigma^2_{Y_1}(x)\}$ and $Y_0|X = x \sim \mathcal{N}\{\mu_{Y_0}(x), \sigma^2_{Y_0}(x)\}$, i.e. we assume that the outcome of interest follows a normal distribution regardless of whether it receives the treatment or not. However, the parameters of the normal distribution change with treatment, leading to $\vartheta_{Y_1}(x) = (\mu_{Y_1}(x), \sigma^2_{Y_1}(x))$ under treatment and $\vartheta_{Y_0}(x) = (\mu_{Y_0}(x), \sigma^2_{Y_0}(x))$ without treatment. The treatment effect on the mean (given characteristics x) is then given by

$$\text{TE}_\mu(x) = \mu_{Y_1}(x) - \mu_{Y_0}(x). \tag{2}$$

By analogy, the treatment effect on the standard deviation (given characteristics x) is

$$\text{TE}_\sigma(x) = \sigma_{Y_1}(x) - \sigma_{Y_0}(x),$$

but we can also easily derive a treatment effect on, for example, the coefficient of variation as $\sigma_{Y_1}(x)/\mu_{Y_1}(x) - \sigma_{Y_0}(x)/\mu_{Y_0}(x)$ or on the quantiles as in equation (1) by evaluating the inverse cumulative distribution functions of the corresponding normal distributions. As a consequence, our parametric distributional approach does not provide one single treatment effect but rather a variety of treatment effects on various distributional features that can be derived from the parameters of the outcome distribution with and without treatment. This is particularly so when replacing the normal distribution with more general types of distributions as formalized in generalized additive models for location, scale and shape (GAMLSSs) (Rigby and Stasinopoulos, 2005). The GAMLSS class is a highly flexible model class that allows all parameters of a conditional distribution to vary with covariates and for non-linear relationships between covariates and predictors. It explicitly and parsimoniously models the distribution of the outcome making a GAMLSS more flexible than a linear model. The main advantage over non-parametric models is that conditioning on covariates is inherent in the framework and thus straightforward.

When moving from binary to continuous treatments, the basic set-up that has been discussed so far remains the same but one must specifically determine the status of the treatment variable D before treatment ($D=d_0$) and after treatment ($D=d_1$). For non-linear models focusing on distributional features beyond the mean, the treatment effect then usually explicitly depends on both d_0 and d_1, i.e.

$$\mathrm{TE}_{\vartheta}(x, d_0, d_1) = \vartheta(x, d_1) - \vartheta(x, d_0),$$

where $\vartheta(x, d)$ represents some distributional quantity given characteristics $X=x$ and treatment status $D=d$. For the original treatment status d_0, one often considers the empirical mean from a sample or some representative values of interest. The change in treatment can also be determined differently, e.g. by changing by 1 unit corresponding to the notion of marginal effects. These marginal effects can be calculated at means (marginal effects for means (MEMs)) or at other representative values of covariates, or as average marginal effects (AMEs). Both MEMs and AMEs can then be formulated for different quantities of the distribution. For example, MEMs can be written as

$$\begin{aligned} \text{MEM on mean} &= E\{Y_i(\bar{d}_0+1)|X=\bar{x}\} - E\{Y_i(\bar{d}_0)|X=\bar{x}\}, \\ \text{AME on mean} &= E\{Y_i(d_{0,i}+1)|X=x_i\} - E\{Y_i(d_{0,i})|X=x_i\}, \end{aligned}$$

with i indexing the individual and $Y_i(d)$ denoting the outcome for individual i given treatment status $D=d$. Instead of a 1-unit change, changes by 1 standard deviation can also be considered for continuous treatments. For a binary treatment, a marginal effect implies a change in the treatment variable from 0 to 1. Thus, if the sample of individuals is representative for the population, AMEs and MEMs for a binary treatment correspond to the estimated ATE or the estimated conditional ATE depending on whether the covariates are fixed at their observed values or at the mean. For continuous treatments the equivalence between marginal effects and the ATE usually does not hold since the treatment often changes from different baseline levels or by different amounts for each individual.

In the case of perfect randomization and compliance, all treatment effects discussed so far could easily be evaluated by including the treatment variable as an additional covariate in a regression analysis. However, even in randomized control trials (RCTs), compliance of the treatment is often not perfectly observed, calling for an instrumental variable (IV) approach. In other settings, where randomization is not possible, the treatment is biased because of self-selection or other sources of endogeneity. Hence, in many quasi-experimental settings, experimental settings with low compliance or when an explanatory or treatment variable is suspected to be endogenous, an IV can still determine a causal effect. An IV is a variable that affects the treatment or an endogenous covariate but not the outcome and therefore provides information on the causal variation in the response of interest induced by the treatment.

Traditional IV estimators are the Wald estimator or the two-stage least squares estimator 2SLS, where, in the first step, the IV is regressed on the treatment variable (and other covariates) via ordinary least squares and the fitted values from this regression replace then the endogenous treatment variable in the regression specification for the variable of main interest. In a more general setting, 2SLS can not only be applied for treatment effects but for any endogenous covariate, i.e. to determine the causal effect of a covariate that is correlated with the error term. Following Imbens and Angrist (1994) and Angrist *et al.* (1996), such an IV analysis recovers only the local ATE of a certain subgroup (the so-called *compliers*, i.e. it enables correction for deviations from perfect randomization but not for deviations from perfect compliance). In contrast, the two-stage residual inclusion estimator 2SRI includes residuals from the first stage of the IV regression instead of the predicted values and in this way enables the ATE to be targeted instead of the local ATE (Basu *et al.*, 2018). The idea of 2SLS is to use only the treatment part that is independent of the unmeasured confounders to explain the outcome, whereas 2SRI splits the unmeasured confounders into a part that is correlated with the treatment and a part that is not (Guo and Small, 2016). Terza *et al.* (2008) reported good performance of 2SRI when the response variable of main interest does not follow a Gaussian distribution and the expectation of the outcome is related to the covariates by means of a non-linear function. In a way, 2SRI is a form of the control function approach (see Wooldridge (2015) for a review) and was applied to generalized additive models (GAMs) by Marra and Radice (2011). The 2SRI estimator has recently become popular within the field of survival analysis, since the Cox model's hazard rate is connected via a non-linear function to the predictor, making it a directly comparable case with Terza *et al.* (2008). It is also in this strand of literature, where the asymptotic theory for 2SRI has been developed (Jiang *et al.*, 2018; Ying *et al.*, 2019).

We draw on the literature on 2SRI and place it within the GAMLSS framework, not only to estimate treatment effects on the conditional mean of the outcome, but on the whole conditional outcome distribution. In this way, we extend the scope of IV regression towards applications dealing with distributional questions that can be consistently answered by using *one* model. To achieve this goal, we propose an IV estimation procedure within the GAMLSS framework, which we call 2SGAMLSS. The purpose of this paper is twofold: we first analyse the performance of our estimator and, second, we demonstrate what additional insights the GAMLSS framework offers when applying 2SGAMLSS to IV regression.

In a simulation study, we assess the ability of 2GAMLSS to estimate the coefficients of the endogenous variable, the MEMs and AMEs on the mean, and the MEMs and AMEs on the standard deviation of which the second two are not captured by previous approaches. We find that our estimator performs particularly well in all non-linear settings as well as in linear settings where the explanatory and endogenous variables are continuous.

We apply our method to a study on electrification in the South African province of KwaZulu-Natal by Dinkelman (2011) that is presented as a motivating example in Section 2. The *ex post* effect of large infrastructural projects such as electrification can often only be estimated by using IVs, making it a relevant example for the method proposed. The study by Dinkelman (2011) analyses the causal effect of rural electrification on employment rates by using the land gradient as the IV to account for the effect that entering the electrification programme was not at random. Using 2SGAMLSS, we account for non-normal outcomes and non-linearities between treatment and instrument, as well as for the neighbourhood structure between administrative units, and we evaluate the effect of electrification on the whole conditional distribution of employment.

We find that the allocation of an electrification project leads to positive marginal effects on the mean (AMEs and MEMs) for employment rates of females, negative effects for employment of males, and a reduced conditional standard deviation for both, indicating a homogenization in employment rates. However, these effects are not statistically significant.

The remainder of this paper is structured as follows: Section 2 presents the electrification study and data used, whereas Section 3 briefly reviews existing non-linear IV approaches and introduces 2SGAMLSS. Section 4 performs an extensive simulation study whereas the application on rural electrification is presented in Section 5. Finally, Section 6 concludes.

2. Motivating example

The importance of access to electricity for everyone has gained considerable attention from the international community and is highlighted in the sustainable development goals that were set by the United Nations General Assembly for 2030. Assessing the direct effects of access to electricity on both the individual and the aggregate level is crucial to design electricity programmes to improve livelihoods. Studies show that access to electricity provides positive effects on labour productivity (Lipscomb *et al.*, 2013), household consumption (van de Walle *et al.*, 2017) and individual access to jobs (Grogan and Sadanand, 2013), among others. Given the nature of electricity installations being related to natural settings and political decision making, deriving a causal estimate is difficult. Many studies rely on IV techniques to disentangle such an effect, making electrification an ideal topic for our proposed method.

To apply 2SGAMLSS and to demonstrate what additional information we can draw from it, we rely on rural electrification data from South Africa and replicate a study by Dinkelman (2011) by using the proposed 2SGAMLSS. Using an IV strategy, Dinkelman (2011) estimated the effect of electrification on employment rates for females and males in rural KwaZulu-Natal communities. After Apartheid, during which many households were denied access to electricity, South Africa's electricity utility (Eskom) committed to supplying access to electrification for everyone from 1995 onwards. The following electrification roll-out is considered to suffer from selection bias since flourishing or politically important areas were presumably targeted first. Hence, Dinkelman used in her main analyses an IV strategy with land gradient as the instrument for the allocation of electrification. The idea is that the land gradient is related to project allocation to communities since a higher gradient increases the costs of electrification but is unrelated to the labour market outcomes, which she showed in a placebo experiment.

The data set in the original work is a combination of two census surveys: administrative data on the roll-out and geographical data. The data that we use to replicate her IV analysis are aggregated at the community level and were collected in two waves; one in 1996 (which is used as baseline), and the other in 2001. For our analysis we consider two response variables: the difference in employment between 1996 and 2001 for male and female individuals. The responses, which are denoted by Δ_t prop_male_emp and Δ_t prop_female_emp, are created by taking the proportion of males or females employed with respect to the population in 2001 minus the baseline proportion of 1996.

Table 1 displays the variables that were considered in both the original and our analysis. The endogenous variable (treatment variable) Eskom is a binary indicator that is equal to 1 if a community received an electrification project between 1996 and 2001, and 0 otherwise. The endogenous treatment covers 20% of the $N = 1816$ communities in the sample. The IV Gradient indicates the average land inclination of each community in degrees. Fig. 1 summarizes the example's settings. The outcome variables of interest are the differences in employment rates that are suspected to be influenced by the Eskom electrification programme. Participation in the

Table 1. Summary statistics for the baseline covariates in the analysis†

Variable	*Description*	*Mean*	*Standard deviation*
Δ_t prop_female_emp	Difference in proportion of employment of females	−0.00	0.07
Δ_t prop_male_emp	Difference in proportion of employment of males	−0.04	0.09
Eskom	Electrification project allocation	0.20	0.40
Gradient	Mean land gradient or inclination	10.10	4.89
prop_hh_fem	Proportion of female-led households	0.55	0.13
hh_povrate	Poverty rate	0.61	0.19
sexratio	Sex ratio $N_{\text{females}}/N_{\text{males}}$	1.48	0.28
prop_indianwhite	Proportion of Indian or white adults	0.00	0.01
kms_to_road	Distance (km) to road	37.95	24.57
kms_to_town	Distance (km) to town	38.57	18.12
kms_to_grid	Distance (km) from grid	19.06	13.32
hh_density	Household density	22.05	30.48
prop_hs_male	Proportion of men with high school education	0.06	0.05
prop_hs_fem	Proportion of women with high school education	0.07	0.05
d_prop_flush	Difference in toilet access	0.03	0.08
d_prop_water	Difference in water access	0.01	0.26

†Number of communities $N=1816$. Number of districts $G=10$.

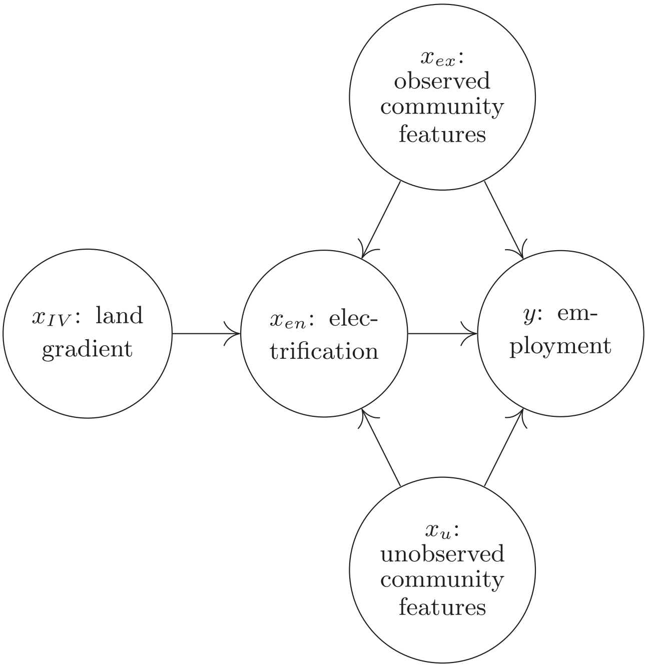

x_{ex}: observed community features
x_{IV}: land gradient
x_{en}: electrification
y: employment
x_u: unobserved community features

Fig. 1. IV setting for the Eskom programme

programme is prone to endogeneity due to some areas being of higher political interest and is thus instrumented by land gradient.

2SGAMLSS augments the electrification analysis in four ways.

(a) Instead of assuming a linear relationship between the instrument (or other regressors) and the endogenous Eskom treatment, this relationship is modelled flexibly.
(b) In addition to assuming a normal distribution for the employment outcome, we employ a logistic distribution. The *qq*-plots in Fig. B3 in the on-line appendix suggest a slightly better fit of the logistic distribution.
(c) Instead of analysing only treatment effects on the mean, we extend the analysis to causal effects on the standard deviation of each outcome variable.
(d) Instead of modelling the 10 districts in KwaZulu-Natal as fixed effects, we account for the neighbouring structure and employ spatial effects via Gaussian Markov random fields.

3. Methodology

3.1. Non-linear instrumental variable regression

The response variable in our application, the differences in employment rates, is possibly non-normally distributed. When considering non-Gaussian outcomes in the context of generalized linear models, the expectation of the outcome is connected to a linear predictor via a one-to-one response function

$$\mathbb{E}(\mathbf{y}|\mathbf{x}_{\mathrm{en}},\mathbf{X}_{\mathrm{ex}},\mathbf{X}_{\mathrm{u}})=h(\mathbf{x}_{\mathrm{en}}\beta_{\mathrm{en}}+\mathbf{X}_{\mathrm{ex}}\boldsymbol{\beta}_{\mathrm{ex}}+\mathbf{X}_{\mathrm{u}}\boldsymbol{\beta}_{\mathrm{u}}),$$

where $\mathbf{y}$ is the outcome variable, $\mathbf{X}_{\mathrm{ex}}$ is an $n \times W_{\mathrm{ex}}$ matrix of exogenous variables, $\mathbf{x}_{\mathrm{en}}$ is a column vector and denotes the endogenous treatment variable and $\mathbf{X}_{\mathrm{u}}$ is an $n \times W_{\mathrm{u}}$ matrix of unobservable confounders. In case the model includes several treatment variables, $\mathbf{x}_{\mathrm{en}}$ is replaced by an $n \times W_{\mathrm{en}}$ matrix of endogenous variables $\mathbf{X}_{\mathrm{en}}$. The corresponding unknown regression coefficients are $\boldsymbol{\beta}_{\mathrm{ex}}$ of dimension $W_{\mathrm{ex}} \times 1$, β_{en}, and $\boldsymbol{\beta}_{\mathrm{u}}$ of dimension $W_{\mathrm{u}} \times 1$. The inverse of the response function $h(\cdot)$ is the link function $g(\cdot)=h^{-1}(\cdot)$. For the remainder of this section we assume that $W_{\mathrm{en}}=1$, i.e. we have only one endogenous variable which is a treatment variable in our example. The reduced form equation of the endogenous explanatory variable can be formulated as

$$\mathbf{x}_{\mathrm{en}}=h_{[1]}(\mathbf{X}_{\mathrm{ex}}\boldsymbol{\delta}_{\mathrm{ex}}+\mathbf{X}_{\mathrm{IV}}\boldsymbol{\delta}_{\mathrm{IV}})+\boldsymbol{\xi},$$

where $h(\cdot)$ is the conditional expectation of $\mathbf{x}_{\mathrm{en}}$ given the exogenous regressors and IVs, and the subscript '[1]' indicates the first-stage model. The matrix $\mathbf{X}_{\mathrm{IV}}$ is of dimension $n \times W_{\mathrm{IV}}$ and contains the instruments. The exogenous regressors are contained within $\mathbf{X}_{\mathrm{ex}}$. The vectors $\boldsymbol{\delta}_{\mathrm{ex}}$ and $\boldsymbol{\delta}_{\mathrm{IV}}$ are of dimension $W_{\mathrm{ex}} \times 1$ and $W_{\mathrm{IV}} \times 1$ respectively, and they contain the unknown first-stage regression coefficients. The number of elements in $\mathbf{X}_{\mathrm{IV}}$ must be equal to or greater than the number of endogenous regressors and $W_{\mathrm{IV}} \geqslant 1$. The term $\boldsymbol{\xi}$ is a vector of errors of dimension $n \times 1$ that contains information about the unobserved confounders. Replacing the endogenous explanatory variable by its ordinary least squares fitted values no longer isolates the exogenous variation in $\mathbf{x}_{\mathrm{en}}$ from the variable that is generated by $\mathbf{x}_{\mathrm{u}}$. To retrieve the effect of an endogenous explanatory variable on the response in a non-linear context, Terza *et al.* (2008) proposed the following procedure called 2SRI.

(a) Obtain the estimates $\hat{\boldsymbol{\delta}}_{\mathrm{ex}}$ and $\hat{\boldsymbol{\delta}}_{\mathrm{IV}}$ from the first-stage regression by using a generalized linear model algorithm. Define the (pseudo)response residuals as

$$\hat{\boldsymbol{\xi}}=\mathbf{x}_{\mathrm{en}}-h_{[1]}(\mathbf{X}_{\mathrm{ex}}\hat{\boldsymbol{\delta}}_{\mathrm{ex}}+\mathbf{X}_{\mathrm{IV}}\hat{\boldsymbol{\delta}}_{\mathrm{IV}}).$$

(b) For the second-stage model, attach the residuals $\hat{\boldsymbol{\xi}}$ as an additional explanatory variable

$$\mathbb{E}(\mathbf{y}|\mathbf{x}_{\text{en}}, \mathbf{X}_{\text{ex}}, \hat{\boldsymbol{\xi}}) = h_{[2]}(\mathbf{x}_{\text{en}}\boldsymbol{\beta}_{\text{en}} + \mathbf{X}_{\text{ex}}\boldsymbol{\beta}_{\text{ex}} + \hat{\boldsymbol{\xi}}\beta_{\hat{\xi}})$$

and estimate the unknown coefficients β_{en} and $\boldsymbol{\beta}_{\text{ex}}$, and the coefficient $\beta_{\hat{\xi}}$ via a generalized linear model or other non-linear method.

The estimated residuals $\hat{\boldsymbol{\xi}}$ will then contain information on the unmeasured confounders. However, the regression coefficient $\beta_{\hat{\xi}}$ cannot be employed to explain the effect of the unobserved confounders on the response, since the variation in $\hat{\boldsymbol{\xi}}$ cannot be assigned to any *meaningful* regressor in particular. This is not problematic, since only accounting for $\mathbf{x}_{\text{u}}$'s absence is necessary to obtain a consistent estimate of β_{en}.

The two-stage GAM procedure 2SGAM that was proposed by Marra and Radice (2011) uses the same approach but relaxes the assumption of strictly linear covariate effects in the first and second stage and relates the dependent variable in both stages to an additive predictor (details on the additive predictor are given in the next subsection). The response residuals from the first stage $\hat{\boldsymbol{\xi}}$ enter the second stage as an additional continuous explanatory variable modelled via smooth functions $f_{\hat{\xi}}$ such that

$$\mathbb{E}(\mathbf{y}|\mathbf{x}_{\text{en}}, \mathbf{X}_{\text{ex}}, \hat{\boldsymbol{\xi}}) = h_{[2]}\left\{\mathbf{X}^*_{\text{ex}}\boldsymbol{\beta}^*_{\text{ex}} + \sum_{l=1}^{L} f_l(\mathbf{x}^+_l) + f_{\hat{\xi}}(\hat{\boldsymbol{\xi}})\right\},$$

where the column vectors of $\mathbf{X}^+ = (\mathbf{X}^+_{\text{ex}}, \mathbf{x}^+_{\text{en}})$ and the residuals are modelled as smooth functions and $\mathbf{X}^*_{\text{ex}}$ as linear effects. The model is estimated by using any GAM method, e.g. via `mgcv` in R (Wood, 2017). The smooth estimates of the first-stage residuals account for the influence of the unmeasured confounders; hence we can consistently estimate the effect of the endogenous explanatory variable. Extending this framework in the presence of multiple endogenous regressors results in a total of $W_{\text{en}} > 1$ first-stage regressions. This produces W_{en} vectors of residuals $\hat{\boldsymbol{\xi}}$ that must be modelled either as linear effects via the regression coefficients $\boldsymbol{\beta}_{\hat{\xi}}$ by using 2SRI or as smooth functions by using 2SGAM.

3.2. Generalized additive models for location, scale and shape

Since the response variable follows a certain distribution, we can move away from considering mere mean effects and shift our interest onto the effect on the whole conditional distribution. The GAMLSS method assumes 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. In the GAMLSS framework, we can specify an equation for each of these parameters of the form

$$g_k(\vartheta_{ik}) = \eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\mathbf{x}_{1i}) + \ldots + f_{J_k}^{\vartheta_k}(\mathbf{x}_{J_k i}), \qquad (3)$$

where the link function g_k ensures compliance with the parameter space and enables modelling a non-linear relationship between the parameter and the predictor η on the right-hand side of equation (3). The predictor $\eta_i^{\vartheta_k}$ has a structured additive form with $\beta_0^{\vartheta_k}$ denoting the overall level of the predictor and functions $f_j^{\vartheta_k}(\mathbf{x}_{ji})$, $j = 1, \ldots, J_k$, can be chosen to model a range of effects of a vector of explanatory variables $\mathbf{x}_{ji}$.

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

(b) Non-linear effects for continuous explanatory variables are captured by smooth functions

$f_j^{\vartheta_k}(\mathbf{x}_{ji}) = f_j^{\vartheta_k}(x_{ji})$ where x_{ji} is a scalar. One way of doing this is by using penalized splines (Eilers and Marx, 1996).

(c) Spatial information can be included via $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = f_j^{\vartheta_k}(s_i)$, where s_i is some spatial information such as geographical co-ordinates or administrative units.

(d) For clustered data, 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 GAMLSS method has the advantage that it estimates the effects on all parameters of a conditional response distribution that can take basically any parametric form and is thus not bounded to the exponential family. Model estimation can be done by maximum likelihood (Rigby and Stasinopoulos, 2005) or Bayesian methods (Klein *et al.*, 2015). Related R packages are gamlss (Stasinopoulos *et al.*, 2017), GJRM (Marra and Radice, 2019) and bamlss (Umlauf *et al.*, 2018).

3.3. Two-stage generalized additive models for location, scale and shape in instrumental variable regression (2SGAMLSS)

We propose the two-stage GAMLSS method 2SGAMLSS: a procedure that in the first stage performs a distributional regression on the reduced form equation of the endogenous variable x_{en}:

$$g_{\text{en}}(\vartheta_{i,\text{en}}) = \eta_i^{\vartheta_{\text{en}}} = \beta_{0,[1]}^{\vartheta_{\text{en}}} + f_{1,[1]}^{\vartheta_{\text{en}}}(\mathbf{x}_{\text{IV},i}) + f_{2,[1]}^{\vartheta_{\text{en}}}(\mathbf{x}_{1i}) + \ldots + f_{J,[1]}^{\vartheta_{\text{en}}}(\mathbf{x}_{J,i}), \tag{4}$$

where $\eta^{\vartheta_{\text{en}}}$ is the structured additive predictor of the conditional expectation of x_{en}, and $g(\cdot) = h^{-1}(\cdot)$ is the link function. The subscript '[1]' indicates that the terms that are specified in equation (4) belong to the first-stage model. The structured additive predictor contains an overall level, as well as effects for the instrument and the remaining exogenous regressors $\mathbf{x}_{1i}, \ldots, \mathbf{x}_{Ji}$. For notational convenience, the subscript k is dropped in equation (4), i.e. a structured additive predictor can be specified for each parameter of the endogenous regressor's distribution. After estimating the regression coefficients in the first-stage model, the conditional expectation of the endogenous regressor and the residuals are computed:

$$\hat{\xi}_i = \mathbf{x}_{\text{en},i} - \mathbb{E}(\mathbf{x}_{\text{en},i} | \hat{\vartheta}_{i,\text{en},1}, \ldots, \hat{\vartheta}_{i,\text{en},K}).$$

Subsequently, all K parameters of the response's density $p(y_i | \vartheta_{i,1}, \ldots, \vartheta_{i,K})$ are regressed on the explanatory variables and the residuals:

$$g_k(\vartheta_{i,k}) = \eta_i^{\vartheta_k} = \beta_{0,[2]}^{\vartheta_k} + f_{1,[2]}^{\vartheta_k}(\mathbf{x}_{1i}) + \ldots + f_{J_k,[2]}^{\vartheta_k}(\mathbf{x}_{J_k i}) + f_{\hat{\xi},[2]}^{\vartheta_k}(\hat{\xi}_i). \tag{5}$$

Here the subscript '[2]' indicates that the components of the K distribution parameters belong to the second-stage model. Note that extending this framework to multiple endogenous regressors results in multiple first-stage models, and having all first-stage residuals attached to the structured additive predictors of the K response distribution parameters. Our proposed procedure resembles that of Marra and Radice (2011) but enables greater flexibility and response distributions that are not members of the exponential family, e.g. zero-inflated distributions.

Especially in an IV setting and in treatment effect evaluation in general, interest often lies in heterogeneous effects. Interaction terms and random coefficients accounting for heterogeneity can be easily included in the second stage. In addition, the GAMLSS method has another notion of heterogeneous effects since they are interpreted conditionally on covariates. One can easily derive AMEs, MEMs or marginal effects at representative values not only for the conditional mean but also for all parameters of the response distribution or other distributional quantities, e.g. the coefficient of variation.

3.4. Confidence intervals

Since 2SGAMLSS relies on two-step estimation, a naive calculation yields intervals that do not necessarily cover their claimed nominal probability, i.e. they will be too narrow. This is because the second-stage regression does not take into account the uncertainty from the quantities that are estimated in the first-stage regression. To represent the uncertainty in the estimated coefficients reliably and to avoid poor coverage, an additional correction must be considered.

Predecessors of 2SGAMLSS have employed a bootstrap pointwise confidence interval correction to restore nominal coverage probabilities. The low coverage probabilities are rectified by employing the joint asymptotic distribution of the GAMLSS maximum likelihood estimators (Stasinopoulos and Rigby, 2007):

$$f(\hat{\boldsymbol{\beta}}|\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where the vector $\hat{\boldsymbol{\beta}}$ contains the estimates of the unknown regression coefficients $\hat{\boldsymbol{\beta}}$, e.g. obtained via the R routine `GJRM::gamlss()`. The algorithm for obtaining confidence intervals is as follows.

(a) Estimate the first-stage model. Draw a total of N_{b} random vectors from a multivariate Gaussian distribution: $\mathcal{N}(\hat{\boldsymbol{\beta}}_{[1]}, \hat{\boldsymbol{\Sigma}}_{[1]})$. Calculate all N_{b} vectors of predictions $\hat{\mathbf{x}}^*_{\mathrm{en},1}, \ldots, \hat{\mathbf{x}}^*_{\mathrm{en},N_{\mathrm{b}}}$, and their respective residuals $\hat{\boldsymbol{\xi}}^*_1, \ldots, \hat{\boldsymbol{\xi}}^*_{N_{\mathrm{b}}}$.

(b) Fit the second-stage model N_{b} times by using the original data attaching the rth vector of residuals. Obtain $\hat{\boldsymbol{\beta}}_{[2],r}$ and $\hat{\boldsymbol{\Sigma}}_{[2],r}$. For each $r = 1, \ldots, N_{\mathrm{b}}$, draw N_{d} random vectors from a multivariate Gaussian distribution, i.e. $\mathcal{N}(\hat{\boldsymbol{\beta}}_{[2],l}, \hat{\boldsymbol{\Sigma}}_{[2],l})$ for $l = 1, \ldots, N_{\mathrm{d}}$.

(c) Calculate the $N_{\mathrm{b}}N_{\mathrm{d}}$ fitted values, e.g. $\hat{f}(\mathbf{x}_{\mathrm{en}})$, and compute the pointwise bootstrap percentile intervals.

Using this procedure, the uncertainty in the residuals $\hat{\boldsymbol{\xi}}$ is accounted for in each of the estimated distribution parameters. We employ the following bootstrap replications for our distributional regression approach: $N_{\mathrm{b}} = N_{\mathrm{d}} = 100$.

4. Simulation study

4.1. Simulation set-up

We investigate the pointwise precision of our proposed 2SGAMLSS estimation procedure in a setting that resembles our considered application, i.e. we fit all estimators by assuming a logistic response distribution and binary endogenous treatment variable. For a more detailed description of the data-generating process (DGP), as well as eight alternative scenarios (S1–S8) using different distributions for the response as well as continuous endogenous treatment variable, see the on-line appendix.

We generate a binary endogenous treatment variable by using a structured additive predictor that consists of effects from an unobserved confounder x_{u} and an instrument x_{IV}:

$$\eta^{\vartheta_{\mathrm{en}}} = \phi_1 f_d(x_{\mathrm{u}}) + \phi_2 f_d(x_{\mathrm{IV}}).$$

The observation index $i = 1, \ldots, n$ is dropped for notational convenience. The parameters ϕ_1 and ϕ_2 are used to control the strength of the instrument, and the severity of the endogeneity. We specify a strong instrument ($|\rho(x_{\mathrm{en}}, f_d(x_{\mathrm{IV}}))| > 0.4$) and severe endogeneity ($|\rho(f_d(x_{\mathrm{u}}), f_d(x_{\mathrm{en}}))| > 0.5$). The distributional parameter ϑ_{en} is obtained by using a response function; then we sample the endogenous treatment x_{en} from a Bernoulli distribution:

$$\vartheta_{\text{en}} = g_{\text{en}}(\eta^{\vartheta_{\text{en}}})^{-1},$$
$$x_{\text{en}} \sim \text{Ber}(\vartheta_{\text{en}}).$$

Afterwards the additive predictors of the response distribution parameters are created by using effects from x_{en}, x_{u} and some exogenous variables x_{ex}, e.g.

$$\eta^{\vartheta_1} = f_d(x_{\text{ex}_1}) + x_{\text{en}}\beta_{\text{en}} + f_d(x_{\text{u}}),$$
$$\eta^{\vartheta_2} = f_d(x_{\text{ex}_2}) + x_{\text{en}}\beta_{\text{en}} + f_d(x_{\text{u}}).$$

The distributional parameters of the response are obtained by applying the appropriate response function to each predictor. Subsequently, a total of n observations of y are sampled from a logistic distribution:

$$\vartheta_k = g_k(\eta^{\vartheta_k})^{-1},$$
$$y \sim \text{logistic}(\vartheta_1, \vartheta_2).$$

The parameter ϑ_1 corresponds to the mean, whereas the scale parameter ϑ_2 corresponds to a transformation of the variance of the response variable. We created two DGPs by using this framework: one in which the $f_d(\cdot)$ were specified to be strictly linear, and another with $f_d(\cdot)$ as non-linear functions. The specifics of these non-linear functions are detailed in the on-line appendix.

The estimated coefficient of the endogenous variable is compared against 2SLS, 2SRI, 2SGAM, a naive GAMLSS (ignores endogeneity) and full GAMLSS (benchmark, includes the unmeasured confounder) estimators. All the non-linear functions were modelled by using penalized splines. The residuals that were obtained in the first stage of 2SGAMLSS are scaled to have unit variance as recommended in Geraci *et al.* (2016). All estimations were performed in R (R Core Team, 2019).

4.2. Target effects for binary and continuous treatments

For the main setting, we report the median of all estimated endogenous coefficients $\hat{\beta}_{\text{en}}$ on the location and scale parameter. For the remaining scenarios in the on-line appendix with non-linear effects, we focus on pointwise precision quantified by using the root-mean-square error in relation to the true effect of $\mathbf{x}_{\text{en}}$:

$$\text{RMSE}\{\hat{f}(\mathbf{x}_{\text{en}})\} = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}\{f(\mathbf{x}_{\text{en},i}) - \hat{f}(\mathbf{x}_{\text{en},i})\}^2\right]},$$

where $\hat{f}(\cdot)$ is the estimated non-linear function evaluation of $\mathbf{x}_{i,\text{en}}$. Additionally, the bias that is incurred by each model is calculated by using

$$\text{bias}\{\hat{f}(\mathbf{x}_{\text{en}})\} = \frac{1}{N}\sum_{i=1}^{N}|\hat{f}(\mathbf{x}_{\text{en},i}) - f(\mathbf{x}_{\text{en},i})|.$$

We report the bias, mean, median and interquantile range IQR, as well as the root-mean-squared error of all Monte Carlo replications of each DGP. These metrics were obtained by using 1000 Monte Carlo replications for sample sizes $N = 500, 2000, 4000$. The uncertainty that is related to estimates obtained via 2SGAMLSS is calculated via coverage probabilities of the bootstrap confidence intervals of $\mathbf{x}_{\text{en}}$. We employ 200 independent data sets using a non-linear DGP. The coverage probabilities were evaluated at confidence levels $\alpha = (0.01, 0.05, 0.1)$.

In addition to the metrics on the coefficient of the endogenous variable, the relative bias between the true and estimated MEMs and AMEs, both on the mean and on the standard deviation sd can be considered. For binary treatments, the MEMs and AMEs are given by

$$\begin{aligned}
\text{MEM on mean} &= E\{Y_i(1)|X_{\text{ex}}=\bar{x}_{\text{ex}}\} - E\{Y_i(0)|X_{\text{ex}}=\bar{x}_{\text{ex}}\},\\
\text{MEM on sd} &= \text{SD}\{Y_i(1)|X_{\text{ex}}=\bar{x}_{\text{ex}}\} - \text{SD}\{Y_i(0)|X_{\text{ex}}=\bar{x}_{\text{ex}}\},\\
\text{AME on mean} &= E\{Y_i(1)|X_{\text{ex}}=x_{\text{ex},i}\} - E\{Y_i(0)|X_{\text{ex}}=x_{\text{ex},i}\},\\
\text{AME on sd} &= \text{SD}\{Y_i(1)|X_{\text{ex}}=x_{\text{ex},i}\} - \text{SD}\{Y_i(0)|X_{\text{ex}}=x_{\text{ex},i}\},
\end{aligned}$$

Equivalently, for continuous treatments MEMs and AMEs can be calculated by

$$\begin{aligned}
\text{MEM on mean} &= E\{Y_i(\bar{x}_{\text{en}}+\text{sd})|X_{\text{ex}}=\bar{x}_{\text{ex}}\} - E\{Y_i(\bar{x}_{\text{en}})|X_{\text{ex}}=\bar{x}_{\text{ex}}\},\\
\text{MEM on SD} &= \text{SD}\{Y_i(\bar{x}_{\text{en}}+\text{sd})|X_{\text{ex}}=\bar{x}_{\text{ex}}\} - \text{SD}\{Y_i(\bar{x}_{\text{en}})|X_{\text{ex}}=\bar{x}_{\text{ex}}\},\\
\text{AME on mean} &= E\{Y_i(x_{\text{en},i}+\text{sd})|X_{\text{ex}}=x_{\text{ex},i}\} - E\{Y_i(x_{\text{en},i})|X_{\text{ex}}=x_{\text{ex},i}\},\\
\text{AME on SD} &= \text{SD}\{Y_i(x_{\text{en},i}+\text{sd})|X_{\text{ex}}=x_{\text{ex},i}) - \text{SD}\{Y_i(x_{\text{en},i})|X_{\text{ex}}=x_{\text{ex},i}\}.
\end{aligned}$$

In the simulation study, we calculate both AMEs and MEMs for binary treatments and focus on the MEMs for continuous treatments with a change of 1 standard deviation to meet the range of the treatment. Focusing on the marginal effects at means or other values has the advantage that we can consider different settings and scenarios of the treatment change, i.e., hypothetically, we could assign different amounts of the treatment to certain individuals to obtain potential outcomes. The advantage of the AMEs is that they provide an overall measure of the actual individuals in the sample. However, AMEs are not adequate if individuals with a certain covariate combination have a very different effect compared with another individual with different covariate values. In practice, we thus recommend calculating AMEs or MEMs or both and, if there is a specific covariate combination that the researcher is interested in, marginal effects at representative values should also be reported.

4.3. Results

Table 2 shows the median of the estimated coefficient $\hat{\beta}_{\text{en}}$ on the response distribution parameters ϑ_1 and ϑ_2 across all Monte Carlo replications. Coefficients estimated by using 2SGAMLSS consistently match those of the benchmark model regardless of DGP and sample size, i.e. estimation of β_{en} is not affected by other linear or non-linear functional forms of the additional covariates. The observed deviations between the medians of 2SGAMLSS and benchmark estimates are very small on both the location (ϑ_1) and scale (ϑ_2) parameters. The estimates of β_{en} produced by the naive GAMLSS method either underestimate or overestimate the covariate's effect on both distributional parameters. This issue is not corrected by increasing the sample size. Standard IV estimation via 2SLS exhibits noticeable deviations from $\hat{\beta}_{\text{en}}$ estimated by the benchmark model as well as our proposed estimator for small sample sizes. Other non-linear IV methods considered such as 2SRI also tend to underestimate the coefficient of the endogenous treatment variable given small sample sizes; see the columns dedicated to $N=500$ in Table 2. Given larger sample sizes, the estimates from 2SLS, 2SRI, 2SGAM and 2SGAMLSS show minimal differences. However, this behaviour is not observed throughout non-Gaussian responses; see for example the section in the on-line appendix Table A4 dedicated to scenario S6, and S7. Overall, coefficients estimated by using 2SGAMLSS repeatedly match the coefficients delivered by the benchmark model.

The GAMLSS framework enables us to recover the effect of the endogenous regressor on all parameters of the response distribution. The coefficients of x_{en}'s effect on the scale parameter ϑ_2

Table 2. Median $\hat{\beta}_{\text{en}}$ on both distribution parameters ϑ_1 and ϑ_2 of a logistic-distributed response across various sample sizes by using 1000 Monte Carlo replications†

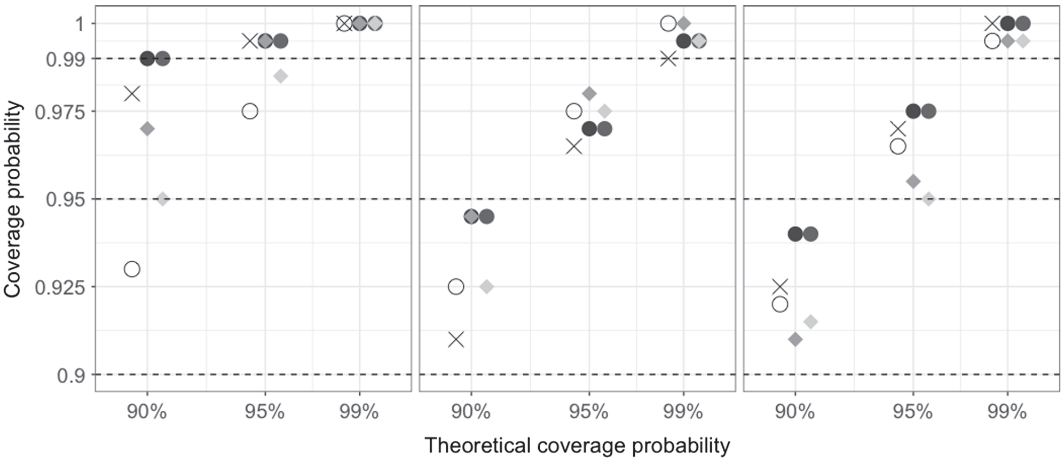
Method	*Results for $N=500$*		*Results for $N=2000$*		*Results for $N=4000$*	
	Linear	*Non-linear*	*Linear*	*Non-linear*	*Linear*	*Non-linear*
Estimated $\hat{\beta}_{\text{en}}$ on ϑ_1						
Naive	1.291	0.674	1.236	0.642	1.277	0.657
2SLS	1.079	0.895	0.923	1.098	1.009	1.068
2SRI	1.028	0.816	0.946	1.098	1.007	1.058
2SGAM	1.125	1.032	0.938	0.954	0.965	1.118
2SGAMLSS	1.188	1.044	0.997	0.994	0.963	1.040
Benchmark	1.099	1.069	0.958	0.976	0.996	0.998
Estimated $\hat{\beta}_{\text{en}}$ on ϑ_2						
Naive	1.277	1.232	1.270	1.221	1.273	1.219
2SGAMLSS	0.984	1.009	0.946	0.969	0.944	0.967
Benchmark	1.007	1.014	0.999	1.003	1.002	1.001

†The procedures 2SLS, 2SRI and 2SGAM are fitted on a Gaussian distribution, since the logistic distribution cannot be fitted by these procedures, i.e. these estimators are misspecified.

from 2SGAMLSS exhibit similar values compared with the benchmark model across DGPs and sample sizes. Similarly to the estimates on the location parameter, naive GAMLSS estimates of β_{en} on ϑ_2 remain biased across multiple sample sizes. The lower section of Table 2 shows how our proposed estimator matches the benchmark estimate on the scale parameter across the sample sizes considered. Such consistent estimation of β_{en} is observed across different response distributions; see Table A5 in the on-line appendix. 2SLS, 2SRI and 2SGAM are limited by a constant scale parameter assumption; therefore no coefficient can be recovered for the scale parameter, or any other potential parameter of the conditional response distribution.

If the endogenous variable considered is continuous (scenarios S1–S4), 2SGAMLSS delivers precise effect estimates for x_{en} across all response distribution parameters. This behaviour is observed on both linear and non-linear DGPs. Tables A2 and A3 (in the on-line appendix) show that the mean and median biases of 2SGAMLSS are considerably lower than either naive estimation and other non-linear IV methods (2SRI and 2SGAM) in the location parameter for samples containing around $N=2000$ observations or more. IQR of the bias of x_{en} also exhibits smaller values for our proposed 2SGAMLSS, indicating a narrower distribution of the bias incurred. For 2SGAMLSS, the values of all metrics considered, i.e. the mean and median bias, the bias's IQR, and RMSE, approach the benchmark values as the sample sizes increases. The metrics on ϑ_2 exhibit overall slightly larger values compared with those for the location parameter ϑ_1, but a trend that favours 2SGAMLSS as the number of observations grows is still noticeable. A somewhat larger sample is essential for 2SGAMLSS to yield consistent effect estimates on the scale parameter. Precise estimation of effects on the scale parameter is crucial since it improves the estimation of heterogeneity in Gaussian responses as in scenario S1 or strictly positive responses as in scenario S4, and overdispersion in count responses as in scenario S3.

Fig. 2 depicts the coverage probabilities of the 2SGAMLSS bootstrap confidence intervals. By setting the bootstrap parameters to $N_{\text{b}}=N_{\text{d}}=100$, the intervals achieve satisfactory cover-



(a) (b) (c)

Fig. 2. Coverage probabilities of the bootstrap confidence intervals for β_{en}, and target treatment effects on the mean and standard deviation at various confidence levels (90%, 95%, 99%) across sample sizes (a) $N=500$, (b) $N=2000$ and (c) $N=40000$ by using 200 Monte Carlo replications: ○, location parameter; ●, AME (mean); ●, MEM (mean); ×, scale parameter; ◆, AME (standard deviation); ◆, MEM (standard deviation)

Table 3. Median relative bias of the estimated MEM†

Method	*Results for N = 500*		*Results for N = 2000*		*Results for N = 4000*	
	Linear	*Non-linear*	*Linear*	*Non-linear*	*Linear*	*Non-linear*
Estimated $\widehat{MEM}$ on the mean						
Naive	−0.291	0.326	−0.236	0.358	−0.277	0.343
2SLS	−0.079	0.105	0.077	−0.098	−0.009	−0.068
2SRI	−0.028	0.184	0.054	−0.098	−0.007	−0.058
2SGAM	−0.125	−0.032	0.062	0.046	0.035	−0.118
2SGAMLSS	−0.180	−0.035	0.003	0.006	0.038	−0.040
Benchmark	−0.099	−0.069	0.042	0.024	0.004	0.002
Estimated $\widehat{MEM}$ on the standard deviation						
Naive	−0.356	−0.414	−0.356	−0.403	−0.358	−0.404
2SGAMLSS	0.013	−0.070	0.033	−0.036	0.029	−0.022
Benchmark	0.004	0.035	0.002	0.041	−0.003	0.043

†The procedures 2SLS, 2SRI and 2SGAM are fitted on a Gaussian distribution, since the logistic distribution cannot be fitted by these procedures, i.e. these estimators are misspecified.

age probabilities of the endogenous treatment effect for the response distribution parameters considered as well as the target treatment effects.

Tables 3 and 4 show the results for the treatment effects on the mean and standard deviation of the outcome, whereas Fig. A2 in the on-line appendix shows boxplots of these. Results for the relative bias in the remaining scenarios are given in the appendix. Neglecting the endogeneity of the treatment variable (naive) leads to considerable bias of both MEMs and AMEs in the mean regardless of the sample size and type of covariate effects (linear or non-linear). In linear

Table 4. Median relative bias of the estimated AME†

Method	*Results for N = 500*		*Results for N = 2000*		*Results N = 4000*	
	Linear	*Non-linear*	*Linear*	*Non-linear*	*Linear*	*Non-linear*
Estimated $\widehat{AME}$ on the mean						
Naive	−0.291	0.326	−0.236	0.358	−0.277	0.343
2SLS	−0.079	0.105	0.077	−0.098	−0.009	−0.068
2SRI	−0.028	0.184	0.054	−0.098	−0.007	−0.058
2SGAM	−0.125	−0.032	0.062	0.046	0.035	−0.118
2SGAMLSS	−0.180	−0.035	0.003	0.006	0.038	−0.040
Benchmark	−0.099	−0.069	0.042	0.024	0.004	0.002
Estimated $\widehat{AME}$ on the standard deviation						
Naive	−0.305	−0.302	−0.301	−0.291	−0.303	−0.293
2SGAMLSS	0.029	−0.012	0.055	0.040	0.053	0.037
Benchmark	−0.000	0.006	0.003	0.004	−0.001	0.003

†The procedures 2SLS, 2SRI and 2SGAM are fitted on a Gaussian distribution, since the logistic distribution cannot be fitted by these procedures, i.e. these estimators are misspecified.

settings, 2SLS outperforms the naive GAMLSS, but in non-linear cases it incurs a sizable bias compared with 2SRI and 2SGAM.

The 2SGAMLSS estimator repeatedly resembles the benchmark estimator in both types of DGP, as well as across the sample sizes considered. As previously mentioned, the GAMLSS framework enables us to derive treatment effects on different distributional quantities of the outcome of interest. The relative bias that is incurred in the AMEs and MEMs on the standard deviation is shown in the lower sections of Table 3 and 4. 2SGAMLSS outperforms the naive estimator at recovering MEMs on the variance in both DGPs and across sample sizes. The naive estimator exhibits the same behaviour as observed in the treatment effects on the mean, i.e. the bias relative to the true value does not benefit from a simpler DGP (linear) or larger sample sizes. As the remaining procedures considered are restricted to estimating the effects on the mean with a constant scale parameter, they are omitted when standard deviation effects are reported.

Although not shown here, we also computed the relative bias for the MEMs and AMEs on the variance of the outcome. The results for the relative bias on the variance qualitatively match those for the standard deviation.

5. Evaluating the effect of rural electrification on employment rates

5.1. Model

To assess how the better performance of 2SGAMLSS in simulation settings translates into a real world scenario, we now come back to the data set on rural electrification in South Africa. We follow the original approach and fit the first-stage distributional regression on the endogenous treatment Eskom by using the instrument Gradient. Regarding the regressors of the first stage, we employ Dinkelman's (2011) most comprehensive specification. This includes community characteristics at the baseline level to control for different growth paths, controls for the different districts and differences in access to water and sanitation. Our approach differs from the original analysis by modelling the covariates and instrument by using penalized splines instead of strictly linear effects, allowing for data-driven estimation of their (potentially) non-linear functional form. The endogenous covariate Eskom is modelled by using a Bernoulli distribution employing the generalized extreme value link function (by default in GJRM) to relate the distribution parameter $\vartheta_1^{\text{Eskom}}$ with the following structured additive predictor:

$$\begin{aligned}\eta_i^{\text{Eskom}} = {} & \beta_{0,[1]}^{\text{Eskom}} + f_{1,[1]}^{\text{Eskom}}(\text{Gradient}_i) + f_{2,[1]}^{\text{Eskom}}(\text{kms\_to\_grid}_i) + f_{3,[1]}^{\text{Eskom}}(\text{kms\_to\_road}_i)\\ & + f_{4,[1]}^{\text{Eskom}}(\text{kms\_to\_town}_i) + f_{5,[1]}^{\text{Eskom}}(\text{hh\_density}_i) + f_{6,[1]}^{\text{Eskom}}(\text{hh\_povrate}_i)\\ & + f_{7,[1]}^{\text{Eskom}}(\text{prop\_hh\_fem}_i) + f_{8,[1]}^{\text{Eskom}}(\text{sexratio}_i) + f_{9,[1]}^{\text{Eskom}}(\text{prop\_indianwhite}_i)\\ & + f_{10,[1]}^{\text{Eskom}}(\text{prop\_hs\_male}_i) + f_{11,[1]}^{\text{Eskom}}(\text{prop\_hs\_fem}_i) + f_{12,[1]}^{\text{Eskom}}(\text{d\_prop\_flush}_i)\\ & + f_{13,[1]}^{\text{Eskom}}(\text{d\_prop\_water}_i) + f_{14,[1]}^{\text{Eskom}}(\text{district}_i).\end{aligned}$$

After estimating the first-stage regression coefficients, we compute the conditional expectation of Eskom and obtain the residuals:

$$\hat{\xi}_i = \text{Eskom}_i - \mathbb{E}(\text{Eskom}_i|\hat{\vartheta}_1^{\text{Eskom}}).$$

The residuals are scaled to have unit variance as in Section 3.1. The quantity $\hat{\xi}$ enters the second-stage predictors as an additional continuous explanatory variable. Our approach further differs from the original study by employing the logistic distribution instead of a Gaussian distribution for the outcomes. For each response separately (i.e. Δ_t prop_female_emp and Δ_t prop_male_emp), we specify a structured additive predictor for the location parameter ϑ_1 with identity link function:

$$\begin{aligned}\eta_i^{\vartheta_1} = {} & \beta_{0,[2]}^{\vartheta_1} + \beta_{1,[2]}^{\vartheta_1}\text{Eskom}_i + f_{2,[2]}^{\vartheta_1}(\hat{\xi}_i) + f_{3,[2]}^{\vartheta_1}(\text{hh\_povrate}_i) + f_{4,[2]}^{\vartheta_1}(\text{hh\_density}_i)\\ & + f_{5,[2]}^{\vartheta_1}(\text{prop\_hh\_fem}_i) + f_{6,[2]}^{\vartheta_1}(\text{prop\_indianwhite}_i) + f_{7,[2]}^{\vartheta_1}(\text{sexratio}_i)\\ & + f_{8,[2]}^{\vartheta_1}(\text{kms\_to\_road}_i) + f_{9,[2]}^{\vartheta_1}(\text{kms\_to\_town}_i) + f_{10,[2]}^{\vartheta_1}(\text{kms\_to\_grid}_i)\\ & + f_{11,[2]}^{\vartheta_1}(\text{prop\_hs\_male}_i) + f_{12,[2]}^{\vartheta_1}(\text{prop\_hs\_fem}_i) + f_{13,[2]}^{\vartheta_1}(\text{d\_prop\_flush}_i)\\ & + f_{14,[2]}^{\vartheta_1}(\text{d\_prop\_water}_i) + f_{15,[2]}^{\vartheta_1}(\text{district}_i).\end{aligned}$$

We specify a structured additive predictor for the scale parameter ϑ_2 with log-link function that features the same set of covariates as the location parameter.

To account better for district heterogeneity, we model the regressor district as a spatial effect by using Markov random fields instead of fixed effects as in the original study. For ϑ_1, the coefficients that are estimated by 2SGAMLSS will reflect the effect of the covariates and residuals on the expectation of the proportion of males/females employment. In the case of the logistic distribution, the conditional variance is a transformation of the scale parameter. The 2SGAMLSS model was fitted using the R package GJRM (Marra and Radice, 2019), whereas 2SLS from the original study was fitted using AER (Kleiber and Zeileis, 2008).

5.2. First-stage results

In the original study the estimated linear effect of Gradient on Eskom project allocation was negative with $\hat{\beta}_{\text{Gradient}} = -0.0077$. The smooth effect that was estimated in the first stage of 2SGAMLSS that is shown in Fig. 3(a) confirms the existence of an inverse relationship between the instrument and the Eskom project allocation. However, values of Gradient between 0° and 10° land inclination barely have an effect on the structured additive predictor of Eskom. Only when land inclination exceeds 10° will the value of η^{Eskom} start to decrease, *ceteris paribus*. Conventional IV methods such as 2SLS are unable to capture nuances like the range where

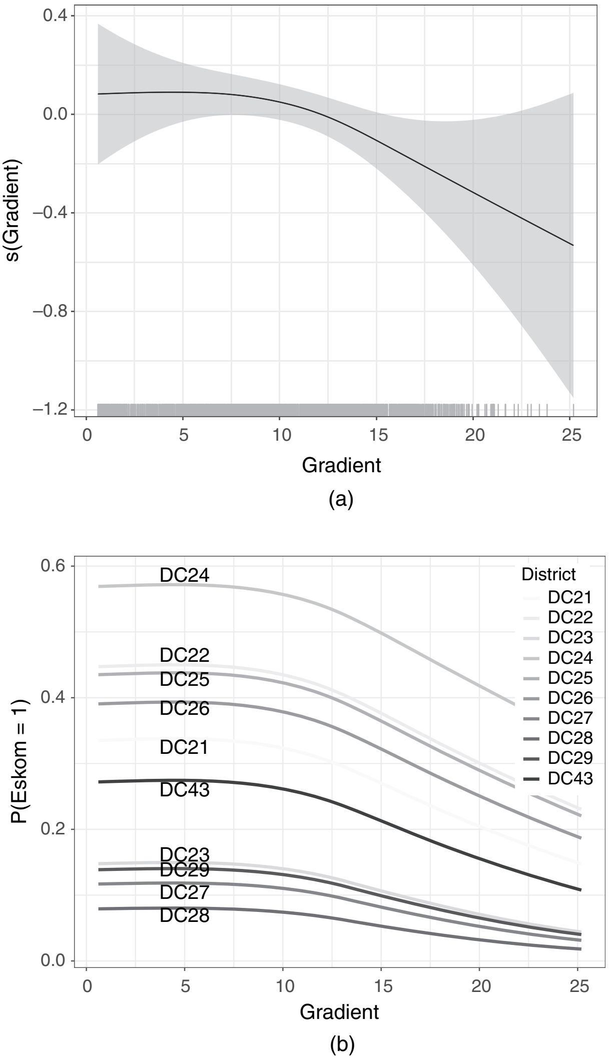


Fig. 3. (a) Estimated smooth effect of the instrument Gradient on the predictor of Eskom with 95% confidence interval and (b) predicted probability of receiving an Eskom project as a function of Gradient across districts DC

Gradient has no effect on the predictor of Eskom. Fig. 3(b) shows the conditional expectation of Eskom as a function of the instrument Gradient, and how the predicted probability of Eskom = 1 varies across the districts of KwaZulu-Natal (see the auxiliary map in Fig. B1 in the on-line appendix to locate neighbouring districts). The inverse relationship between the instrument and Eskom shows steeper descents for some districts (e.g. DC24), but the general trend indicates a decreasing probability of receiving an electrification project as the average land inclination exceeds 10°.

Our proposed approach benefits from the flexible estimation of the instrument's effect, which in turn leads to a better estimate of the first-stage residuals $\hat{\xi}$. The penalized spline estimates of the first-stage residuals on the predictor of ϑ_1 that are shown in Fig. B2 (in the on-line appendix) indicate a strong non-linear effect on the expectation of both response variables. The

Table 5. Regression coefficients (95% confidence intervals), MEMs and AMEs for the electrification data and various IV estimators†

		Results for 2SLS	*Results for 2SGAMLSS*
Δ_t *male employment*			
ϑ_1	Eskom	0.0355	−0.0143
		[−0.0500; 0.2500]	[−0.2615; 0.0156]
	District effects	Fixed	Markov random fields
ϑ_2	Eskom	—	−0.5242
		—	[−1.2302; 0.6032]
	District effects	—	Markov random fields
MEMs or AMEs on mean			−0.0143
			[−0.2615; 0.0156]
MEMs on standard deviation			−0.0010
			[−0.0647; 0.0674]
AMEs on standard deviation			−0.0014
			[−0.2329; 0.0543]
Δ_t *female employment*			
ϑ_1	Eskom	0.0951‡	0.0152
		[0.0500; 0.3000]	[−0.1391; 0.0872]
	District effects	Fixed	Markov random fields
ϑ_2	Eskom	—	−0.8386
		—	[−1.0868; 0.6285]
	District effects	—	Markov random fields
MEMs or AMEs on mean			0.0152
			[−0.1391; 0.0872]
MEMs on standard deviation			−0.0009
			[−0.0523; 0.0499]
AMEs on standard deviation			−0.0012
			[−0.1803, 0.0772]

†All models control for baseline covariates, differences in access to water and sanitation, and district heterogeneity, $N = 1816$. Bootstrap confidence intervals of 2SGAMLSS estimates.
‡$p < 0.1$.

estimates of $\hat{\xi}$ on the structured additive predictor of the scale parameter of both responses also exhibit a non-linear functional form (Fig. B2 in the appendix), which would not have been captured by 2SLS, 2SRI or 2SGAM. It should be noted that these non-linear effects of $\hat{\xi}$ yield no useful interpretation, since the variation in the first-stage residuals cannot be assigned to any explanatory variable. However, the fitted curves validate Dinkelman's (2011) suspicion of the Eskom project's endogeneity.

5.3. Second-stage results

The estimated coefficients for the endogenous treatment in the structured additive predictors of the location and scale parameters of the response distribution are displayed in Table 5. In the original study, the effect of Eskom on both responses was positive and statistically significant for the employment of females. After accounting for possible non-linearities in the covariates, as well as for spatial district heterogeneity, 2SGAMLSS estimates a positive effect of the electricity project allocation on the expectation of the proportion of employment of females. For example, the allocation of an Eskom project will lead to an increase in Δ_t prop_female_emp of 1.52% on average, given that the remaining covariates are held constant. Due to using the identity link and the fact that the location parameter of the logistic distribution equals the conditional mean,

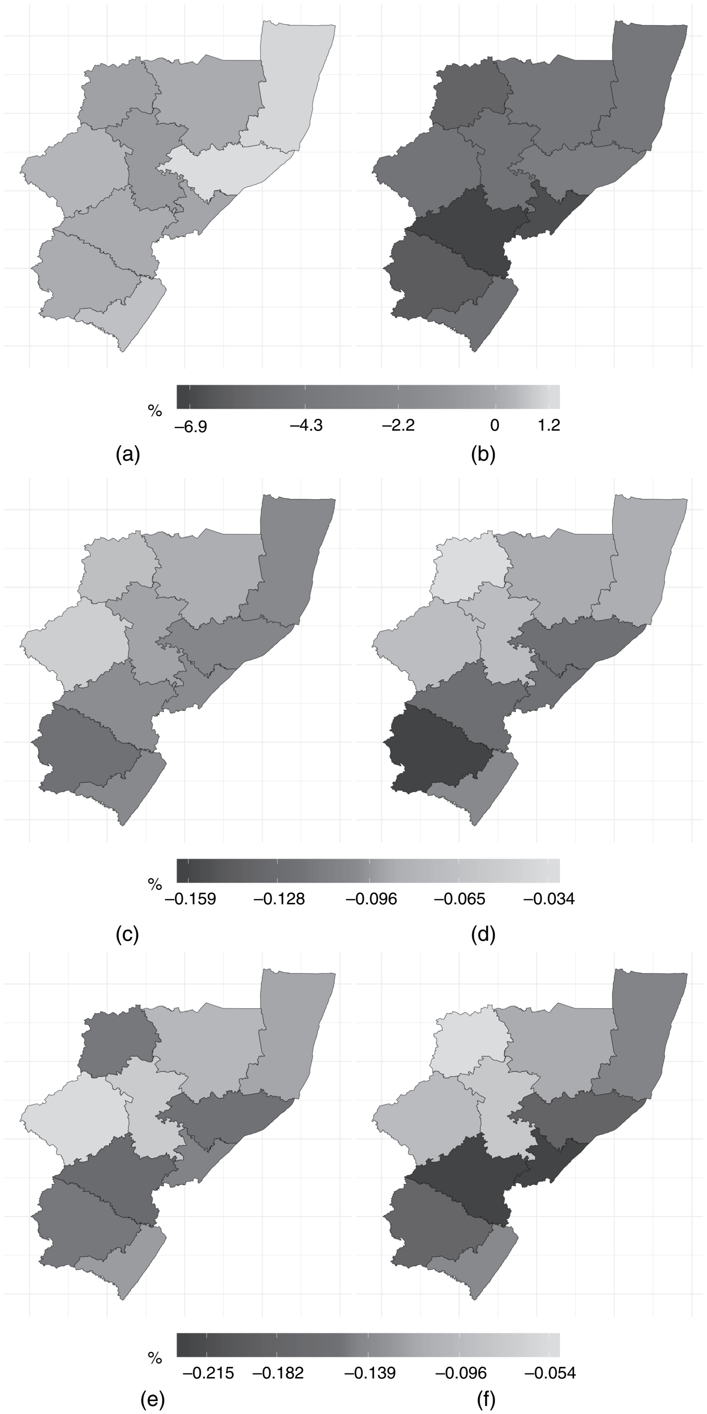


Fig. 4. (a), (b) Estimated MEMs on the mean across districts which correspond to the AMEs for the distribution considered, (c), (d) estimated MEMs on the standard deviation across districts and (e), (f) AMEs on the standard deviation across districts (Δ_t prop_gender_emp$\sim$ logistic(ϑ_1, ϑ_2)): (a), (c), (e) females; (b), (d), (f) males

the effect on ϑ_1 equals the MEMs and AMEs for the mean. The bootstrap confidence intervals indicate that the estimated effect is not significant at the 5% level.

The coefficients for ϑ_2 that were obtained from 2SGAMLSS (logistic) that are shown in Table 5 indicate that the Eskom project allocation has a multiplicative effect on the scale parameter of the response for females of size $\exp(-0.8386) = 0.4323$, given that all other covariates are held constant. Consequently, the MEMs and AMEs for the standard deviation are negative. Note that we decided for the application case to report the effects on the standard deviation due to the scale of the response variable. Variance effects would have been numerically quite small. This means that the conditional standard deviation of the proportion of employment of females for communities that have received an Eskom project is reduced, compared with communities without the electrification project. The coefficient that was estimated by using 2SGAMLSS for Δ_t prop_male_emp suggests that the Eskom project allocation reduces the expected difference in the employment of males by approximately 1.4% and has a multiplicative effect of $\exp(-0.5242) = 0.5920$ on the scale parameter with negative effect on the MEMs and AMEs on the standard deviation.

For a policy maker the combined picture of mean and standard deviation (or variance) effects is of interest. For example, a positive mean effect together with an increase in the standard deviation would have meant that the positive mean effect came mainly through larger benefits for communities that already had higher rates of employment before the programme. Regarding the application, the positive mean and negative standard deviation effect means that a larger increase in employment was experienced by communities that had lower rates of employment before the programme conditionally on covariates. For both men and women, the reduced standard deviation means that the programme led to a homogenization of employment rates between treatment communities compared with control communities, conditionally on covariates. However, the negative mean effect for the employment of males indicates that communities that had higher rates of employment for males before the programme approached the mean rates of employment by experiencing a reduction in rates. Yet, none of the estimated effects for Eskom are significant. The difference between the 2SGAMLSS and the 2SLS estimates for Eskom on the mean of the outcomes could originate from the fact that 2SGAMLSS is based on residual inclusion which tries to recover the ATE, whereas 2SLS estimates the local ATE. Other sources of discrepancy between the fits are 2SGAMLSS's ability to account for possible non-linearities in the covariates' functional form, and the Markov random field representation of the districts' spatial effect.

Fig. 4 depicts the treatment effects for various distributional quantities of both outcomes across the districts of KwaZulu-Natal. The maps for the MEMs and AMEs for the mean indicate that the treatment effect induces a reduction in employment rates for men across all districts. For the employment rates for females, an increase is observed for northern districts and a reduction for a central district. Figs 4(c) and 4(d) display the estimated MEMs for the standard deviation of both outcomes. For women, the MEMs and AMEs for the standard deviation imply a higher degree of homogenization in the east and south than for western districts. For the response for males the estimated treatment effects indicate a reduction in rates of employment accompanied by homogenization of these rates that occurred mostly in the southern districts of KwaZulu-Natal.

6. Concluding remarks

This work proposes an alternative IV estimator which can account for non-normal outcomes, non-linearities between the endogenous variable, instrument and outcome, and can estimate the treatment effect on the whole conditional distribution and not just the mean. The estimator combines a two-step residual inclusion procedure with the GAMLSS method.

A simulation study shows that, especially in non-linear settings, 2SGAMLSS captures well the coefficient of the endogenous variable. Other non-linear IV methods such as 2SRI and 2SGAM show good performance as well but are restricted to estimation of the mean. For linear settings with Gaussian responses, the results of 2SLS and 2SGAMLSS estimation are very similar. Our IV estimator performs best when both the instrument and the endogenous regressor are continuous. In the presence of endogenous binary variables, the endogenous treatment effect estimated by using 2SGAMLSS repeatedly matches a benchmark estimate for all distribution parameters throughout linear and non-linear settings, regardless of the sample size and the response distribution.

We recommend the implementation of 2SGAMLSS in complex IV settings, where the relationships between outcome, instrument and endogenous regressor(s) are *a priori* unknown. In settings, for which some would claim that interest is solely in the mean, we still suggest using 2SGAMLSS for two reasons. Firstly, once we depart from the Gaussian assumption for the response, there are distributions, such as the Gumbel distribution, whose expected mean depends on more than one parameter. There is no reason why one should be dependent on covariates or one should not. Secondly, more on a philosophical side, we argue that most models should involve considerations beyond the mean to answer research questions from multiple perspectives. We follow Rigby *et al.* (2013) and Kneib (2013) who stated that beyond-the-mean considerations are ubiquitous and models dealing with them should not be regarded as an exception. They gave helpful introductions into beyond-the-mean modelling and mentioned various examples that consider the whole conditional outcome distribution. When estimating the effects of a policy programme, even if the primary interest is in the average effect, any analysis should always be concerned with changes in inequality and whether individuals benefit equally.

We replicated and extended an IV study by Dinkelman (2011) who found positive effects of electrification on employment for both female and male individuals. We found that, in an ‘average’ community, the effect on employment is positive but only for the employment of females. The effect of electrification on the employment of males was negative. The endogenous treatment variable also impacts the standard deviation of the conditional response distribution, leading in general to a larger reduction in the standard deviation of the employment of males compared with that of females. These statements regarding the treatment effect on the standard deviation of the conditional outcome distributions complement a proper treatment effect evaluation. Effects on the standard deviation are of interest since, first, any treatment produces not just a mean result but varies around that, and second there is no reason to assume that this variation is equal for all people. In addition to the standard deviation, any other distributional feature such as the Gini coefficient or quantiles can be derived from the results, which is essential when we are concerned about inequality and heterogeneity.

These results are of importance not only for extending the GAMLSS applications to IVs but also to policy makers. Infrastructural projects such as electrification are not only the most cost-intensive projects but also those where treatment effects can often only be consistently estimated by using IVs. The method proposed herein enables more exact estimation of the relationships, improving the guidance and justification for policy makers for those projects.

Acknowledgements

The authors thank two referees and the Associate Editor who kindly reviewed earlier versions of the manuscript and provided very valuable suggestions and comments that substantially improved this work. Maike Hohberg and Thomas Kneib received financial support from the Deutsche Forschungsgemeinschaft within research project KN 922/9-1.

Open access funding enabled and organized by Projekt DEAL.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Basu, A., Coe, N. B. and Chapman, C. G. (2018) 2SLS versus 2SRI: appropriate methods for rare outcomes and/or rare exposures. *Hlth Econ.*, **27**, 937–955.

Dinkelman, T. (2011) The effects of rural electrification on employment: new evidence from South Africa. *Am. Econ. Rev.*, **101**, 3078–3108.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statist. Sci.*, **11**, 89–121.

Geraci, A., Fabbri, D. and Monfardini, C. (2016) Testing exogeneity of multinomial regressors in count data models: does two-stage residual inclusion work? *J. Econmetr. Meth.*, **7**, no. 1.

Grogan, L. and Sadanand, A. (2013) Rural electrification and employment in poor countries: evidence from Nicaragua. *Wrld Devlpmnt*, **43**, 252–265.

Guo, Z. and Small, D. S. (2016) Control function instrumental variable estimation of nonlinear causal effect models. *J. Mach. Learn. Res.*, **17**, 3448–3482.

Imbens, G. W. and Angrist, J. D. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.

Jiang, B., Li, J. and Fine, J. (2018) On two-step residual inclusion estimator for instrument variable additive hazards model. *Biostatist. Epidem.*, **2**, 47–60.

Kleiber, C. and Zeileis, A. (2008) *Applied Econometrics with R*. New York: Springer.

Klein, N., Kneib, T., Lang, S. and Sohn, A. (2015) Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Statist.*, **9**, 1024–1052.

Kneib, T. (2013) Beyond mean regression. *Statist. Modllng*, **13**, 275–303.

Lipscomb, M., Mobarak, A. M. and Barham, T. (2013) Development effects of electrification: evidence from the topographic placement of hydropower plants in Brazil. *Am. Econ. J. Appl. Econ.*, **5**, 200–231.

Marra, G. and Radice, R. (2011) A flexible instrumental variable approach. *Statist. Modllng*, **11**, 581–603.

Marra, G. and Radice, R. (2019) GJRM: generalised joint regression modelling. *R Package Version 0.2*. (Available from `https://CRAN.R-project.org/package=GJRM`.)

Melly, B. and Wüthrich, K. (2017) Local quantile treatment effects. In *Handbook of Quantile Regression* (eds R. Koenker, V. Chernozhukov, X. Hu and L. Peng), pp. 145–164. Boca Raton: Chapman and Hall–CRC.

Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles, section 9 (Engl. transl. D. Dabrowska and T. Speed). *Statist. Sci.*, **5**, 465–472.

R Core Team (2019)*R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.

Rigby, R., Stasinopoulos, D. and Voudouris, V. (2013) A comparison of GAMLSS with quantile regression. *Statist. Modllng*, **13**, 335–348.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Stasinopoulos, M. D. and Rigby, R. A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *J. Statist. Softwr.*, **23**, no. 7, 1–46.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. and De Bastiani, F. (2017) *Flexible Regression and Smoothing: using GAMLSS in R*. Boca Raton: CRC Press.

Terza, J. V., Basu, A. and Rathouz, P. J. (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Hlth Econ.*, **27**, 531–543.

Umlauf, N., Klein, N. and Zeileis, A. (2018) BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *J. Computnl Graph. Statist.*, **27**, 612–627.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*, 4th edn. New York: Springer.

van de Walle, D., Ravallion, M., Mendiratta, V. and Koolwal, G. (2017) Long-term gains from electrification in rural India. *Wrld Bank Econ. Rev.*, **31**, 385–411.

Wood, S. N. (2017) *Generalized Additive Models: an Introduction with R*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Wooldridge, J. M. (2015) Control function methods in applied econometrics. *J. Hum. Resour.*, **50**, 420–445.

Ying, A., Xu, R. and Murphy, J. (2019) Two-stage residual inclusion for survival data and competing risks—an instrumental variable approach with application to SEER-Medicare linked data. *Statist. Med.*, **38**, 1775–1801.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article.