City, University of London Institutional Repository

Modelling the impact of women's education on fertility in Malawi

Luca Zanin

Prometeia

G. Marconi, Bologna 40122, Italy

luca.zanin@studio.unibo.it

Rosalba Radice

Department of Economics, Mathematics and Statistics

Birkbeck, University of London

Malet Street, London WC1E 7HX, U.K.

Giampiero Marra

Department of Statistical Science

University College London

Gower Street, London WC1E 6BT, U.K.

December 17, 2013

Abstract

Many studies have suggested that there is an inverse relationship between education and number of children among women from sub-Saharan Africa countries, including Malawi. However, a crucial limitation of these analyses is that they do not control for the potential endogeneity of education. The aim of our study is to estimate the role of women's education on their number of children in Malawi, accounting for the possible presence of endogeneity and for non-linear effects of continuous observed confounding variables. Our analysis is based on micro data from the 2010 Malawi Demographic Health Survey, by using a flexible instrumental variable regression approach. The results suggest that the relationship of interest is affected by endogeneity and exhibits an inverted-U shape among women who are more than 30 years old and live in rural areas of Malawi, whereas it exhibits an inverse (non-linear) relationship for women living in urban areas.

1

1   Introduction

Malawi is a small country in sub-Saharan Africa that is ranked among the poorest in the world. The population continues to register high birth, infant, and maternal mortality rates (National Statistical Office (NSO) and ICF Macro, 2011; IMF, 2012; MDHS, 2012; Strulik and Vollmer, 2013). However, some progress has been made in Malawi over the last twenty years. For example, the mean number of children born alive over a woman's lifetime declined from 6.7 in 1992 to 5.7 in 2010. In the same period, the under-five mortality rate per 1000 live births decreased from 233.8 to 112.1 (MDHS, 2012). It is plausible that these trends are partly associated with a decline in female illiteracy; the percentage of women with no education decreased from 47.2% to 15.2% between 1992 and 2010 (MDHS, 2012). Education plays a crucial role in individuals' lives because the knowledge accumulated through schooling stimulates new ideas, enhances individual talents, and improves social capital, job opportunities, knowledge about health and the ability to make decisions (e.g., Martin, 1995; Chen and Li, 2009; Bhalotra and Rawlings, 2011).

It is well-documented that women's education is negatively associated with their number of children (e.g., Colclough, 1982; Martin, 1995; Strauss and Thomas, 1995; Kalipeni, 1997; Kravdal, 2000; Basu, 2002; Stephenson *et al.*, 2007; Kazembe, 2009; Luz and Kc, 2011; MDHS, 2012). However, despite existing evidence of a negative correlation between the two variables, the observed association does not imply causation. The omission of unobserved confounding variables (such as abilities, motivation, and preferences) that are strongly associated with education and are factors in a woman's number of children may lead to a biased estimate of the relationship. Therefore, studies that treat education as exogenous by assuming the absence of unobserved confounding effects cannot determine whether schooling causally affects the number of children. To better understand this relationship, researchers should obtain estimates that account for observed and unobserved confounding variables.

Empirical studies that have been conducted using data from Botswana, Indonesia and Nigeria

2

support the presence of unobserved confounders/endogeneity in the relationship between education and number of children (e.g., Angeles *et al.*, 2005; Osili and Long, 2008; Marra and Radice, 2011b; Sobotka *et al.*, 2013). To date, this issue has not been investigated for Malawi. Research shows an inverse relationship between education and number of children in Malawi (Kalipeni, 1997; Kazembe, 2009). It is not clear, however, whether this relationship is affected by endogeneity, which has to be accounted for to ensure that the effect of education is correctly quantified. Additionally, researchers usually account for the confounding effects of observed continuous covariates by including linear, polynomial, or categorical terms in the model (e.g., Bailey, 1989; Kalipeni, 1997; Kravdal, 2000; Kazembe, 2009). But the correct functional form is rarely known *a priori*, and incorrect specifications typically lead to residual confounding which, in turn, leads to inconsistent parameter estimates (e.g., Marra and Radice, 2011a and references therein). Even more problematic is the misspecification of the functional form of the relationship of interest; such misspecification may prevent researchers from identifying interesting dependencies. Both issues, endogeneity and functional form misspecification, call for an approach that can account for unobserved confounders and avoids the downfalls of categorization and/or *a priori* specified parametric models.

Improving our understanding of the relationship between education and number of children contributes to the literature on global health policy-making that focuses on reducing the number of children to more sustainable levels in poor countries. Health policy-makers are interested in knowing if investments in formal education can reduce the number of children and, if so, what mechanisms govern this relationship. In this paper, we provide a refined analysis of the impact of education on the number of children among Malawian women. Our empirical analysis is based on micro data from the 2010 Malawi Demographic Health Survey (MDHS, `www.measuredhs.com`). Specifically, we use a flexible approach, the so-called two-stage generalised additive model (2SGAM), as proposed by Marra and Radice (2011a). This method accounts for both observed and unobserved confounding effects and eliminates the need to specify a priori the analytical form of the treatment-response and confounder-response relationships. This study makes several contributions to the literature. First, we confirm that education has a significant impact on women's number of children in Malawi. Second, the investigated relationship seems to be affected by endogeneity within the sample of women who live in rural area but not for women who live in urban areas.

3

In particular, the fertility-education relationship exhibits an inverted-U shape, especially among women who are more than 30 years old and who live in rural areas. In contrast, we observe an inverse (non-linear) relationship for women living in urban areas. Third, the use of a flexible modelling approach allows us to relax assumptions related to specific functional forms for continuous covariates, thereby minimising specification errors by allowing the data to determine the appropriate relationships (i.e., linear or non-linear).

This article is organised as follows. Section 2 describes the data. Section 3 introduces the model specification and modelling strategy, and Section 4 presents the results of the empirical analysis. Section 5 concludes with a discussion of the findings and with policy recommendations.

2 Material

We investigated the impact of education on number of children in Malawi using MDHS micro data collected between June and November of 2010 by the NSO in partnership with the Ministry of Health Community Sciences Unit (CHSU). The MDHS is part of a programme that is conducted in several developing countries. The MDHS is the fourth survey of its kind to be conducted in Malawi and provides cross-sectional information for scholars and policy-makers on a variety of topics about women's lives, including their background characteristics (i.e., education, residential history, media exposure), number of children, knowledge and use of family planning methods, marriage and sexual activity, nutritional status of mothers and children, use of maternal and child health services, domestic violence, and other relevant variables (NSO and ICF macro, 2011).

The 2010 MDHS survey included respondents from twenty-seven of Malawi's districts that were stratified into urban and rural areas for a total of fifty-four sampling strata. The 2010 MDHS used a two-stage sampling design. In the first stage, a sample of 849 areas (158 urban and 691 rural areas) was chosen with a probability of selection proportional to the size of the areas, which was defined by the number of households in each area. At the second stage, a fixed number of households was selected in each area using equal probability systematic sampling. For further details about the sampling techniques used, we refer the reader to NSO and ICF Macro (2011). The final sample includes 27307 households; 3160 in urban areas and 24185 in rural areas. However, approximately two thousand households were unoccupied. Therefore, respondents from

24825 households were successfully interviewed (response rate of 98%), including 23020 eligible women between the ages of 15 and 49 (response rate of 96.9%).

2.1 Number of children ever born per woman

The variable of interest, `children`, is the number of children ever born to each female respondent. Information for this variable was collected through a specific question included in the section entitled 'Reproduction' of the survey's structured questionnaire (NSO and ICF Macro, 2011 - p. 484). The question reads: '*Now I would like to ask about all the births you have had during your life. Have you ever given birth?*' Possible answers were *yes* or *no*. If respondents replied *yes*, further information about the number of sons and daughters living at home, away from home and those who died after birth was collected. Henceforth, we will use the terms 'number of children' and 'fertility' interchangeably, as done in the literature. The first plot in Figure 1 displays the frequency distribution of `children`, which is skewed to the right, with a mean value of 3.1 children per woman and a standard deviation equal to 2.7. Malawian women at the end of their reproductive lives (45-49 years) have an average of 6.8 children ever born. The total fertility rate (computed by summing age-specific fertility rates) is 5.7 children per woman.

INSERT FIGURE 1 HERE

2.2 Education

Women's education (referred to as `education`) was measured in terms of years of schooling by combining information on educational level attended and the grade attained. Years of schooling is widely used as a proxy variable for education because knowledge and skills should increase with advancement in schooling. However, one limitation of this variable is that it does not measure performance in school, curriculum content, and other qualitative aspects of education (e.g., Zanin and Marra, 2012). Malawi's education system comprises eight years of primary school (referred to as Standards 1 through 8), four years of secondary school (known as Forms 1 to 4), and four years of university education. The official age of entry into primary school is six years old, and the school calendar begins in January, with long holidays in November and December (Kadzamira

and Rose, 2010).

In 1994, the newly elected government of Malawi announced the abolition of primary school fees as a strategy to increase educational opportunity and of furthering its long-term goal of alleviating poverty. However, twenty years after the introduction of this policy, several critical issues remain unresolved. Scholars have highlighted the persistence of three crucial problems in Malawi's educational system: (a) delayed enrolment in primary school for some children, (b) children dropping out before completing their primary education, and (c) the low quality of schooling. These problems are related and represent barriers to the achievement of universal primary education in Malawi (MMDG, 2010).

Moyi (2010) observed that 20% of boys and girls started primary school after the official entry age (8+ years), whereas the remaining 80% of children enrolled on time (<7 years). He observed that the main reasons for delays in enrollment are the distance/walking time to the nearest primary school (especially in rural areas) and the cost of schooling, especially for the poorest households, despite the abolition of fees (Kadzamira and Rose, 2003; Moyi, 2010). In addition, some characteristics of family life in Malawi, such as the need for children to work to help their parents, are among the reasons for delayed enrollments in primary school.

The drop-out rate is another critical issue in Malawi. The second plot in Figure 1 shows that a large portion of women between the ages of 15 and 49 do not complete primary school (i.e., they complete less than 8 years of schooling). The drop-out rate may be influenced by numerous socio-economic and cultural issues that are deeply rooted in the country. Using cross-sectional data, Schafer (2006) found that Malawian children's school drop-out rates are relatively low in households with good education levels and a positive perception of school quality as well as trust in local schools and institutions. In contrast, the probability of a student's dropping out is higher when the student comes from a larger household or a Muslim household. Other factors that affect the drop-out rate are household poverty, delays in enrolment, the cost of schooling, a lack of interest in education, and early marriage (e.g., Kadzamira and Rose, 2003; Chimombo, 2005; Manda and Meyer, 2005; Moyi, 2010). The poor quality of education is another weakness of Malawi's education system. Chimombo (2005) highlighted a series of factors that can compromise the quality of education, including the scarcity of equipment required for effective teaching (e.g., textbooks, desks), poor condition of many schools, and low teacher-student ratio (this ratio ranged

from 1:142 in the Salima school district to 1:53 in the Lilongwe school district). It is important to note that the goal of universal primary education cannot be attained without ensuring school quality because increasing the number of people who complete primary education does not ensure they have the relevant knowledge and skills to make important contributions to the development of the country (Chimombo, 2005). This brief discussion draws the fragility of the education system in Malawi. However, the available micro data do not include information on the quality of the schools in the territory, information on the cultural background of the household of origin of women, or indicators of the preference, abilities, and motivation of women at the time of schooling. Thus, in conducting our analysis, we are not able to consider some relevant factors that influence women's education.

2.3 Observed confounders

The micro data include a number of observed confounders that may affect women's number of children and the relationship between education and fertility.

Region of residence (`region`): this is defined as a categorical predictor that identifies the geographical region of Malawi where each woman was interviewed. Malawi is divided into three regions: northern, central, and southern. Descriptive statistics (Table 1) show that 47.65%, 34.15% and 18.20% of women live in the southern, central and northern regions of the country, respectively. This variable should capture socio-cultural differences between regions that are related to women's fertility.

Urban area (`urban`): in addition to region, we also considered the metropolitan character of respondents' place of residence. We constructed a dichotomous variable that takes the value of one if a woman lives in an urban area and zero otherwise (i.e., the woman lives in a rural area). Several studies show that women who live in urban areas have lower fertility rates than those who live in rural areas (e.g., Kalipeni, 1997; Kravdal, 2003; Sobotka *et al.*, 2013).

Woman's age (`age`): the age of respondents plays a crucial role in socio-demographic analyses because age affects many aspects of life such as education, fertility, marriage, retirement, mortality, and so on (e.g., Palamuleni, 2011). Women's age was coded as a continuous variable and ranged between 15 and 49 years. This age-range is typically considered to span women's reproductive

lives.

Never married women (`never_married`): We created a binary variable which is equal to one if a woman has never married and zero otherwise. Scholars have observed that births usually occur within marriage in developing countries (e.g., Martin, 1995; Manda and Meyer, 2005; Kazembe, 2009). In the present study, we have not considered the woman's age at marriage. Some observational analyses have noted that self-reported data on the year of marriage may suffer from recall errors when these data are provided by older women recalling an event that happened many years ago or from under-reporting when they are provided by young women (e.g., Blanc and Rutenberg, 1990; Harwood-Lejeune, 2000). Thus, as a precaution, in the models described in Section 3.2, we have excluded the woman's age at marriage as an observed confounding factor to minimise the use of variables that may be affected by potential uncontrollable measurement errors (see also Sobotka *et al.*, 2013).

Religion's affiliation (`religion`): we grouped religions in Malawi into four main categories: Catholic, Protestant (including CCAP, Anglican, Eleventh Day Advent/Baptist, and other Christian), Muslim, and other or no religion. Heaton (2011) found that Muslims' fertility is higher than Christians' fertility in many developing countries, whereas the difference between Catholics and Protestants is relatively minor. Our sample is composed mostly of Protestant women (68.72%), whereas Catholic and Muslim women represent 20.29% and 10.99% of Malawian women in the sample, respectively (see Table 1).

<div align="center">**INSERT TABLE 1 HERE**</div>

3 Modelling the fertility-education relationship

In analysing the relationship between women's education and their fertility, it is important to account for the possible presence of unobserved confounding effects and of non-linearities in the relationship between observed confounding effects and education. Faced with these issues, we employed a flexible model, the two-stage generalised additive model (2SGAM) introduced by Marra and Radice (2011a). For comparison, we also employed a classic GAM approach (here referred to as naive model) which can model flexibly covariate effects but cannot account for

unobserved confounders.

3.1 Model representation, estimation and inference

Before applying 2SGAM to the fertility data, we briefly discuss the basic ideas underlying the adopted approach in a general context.

GAMs extend generalized linear models (GLMs) by allowing the determination of possible non-linear effects of predictors on the response variable. A GAM has model structure

$$\mathbf{y} = \mathbf{g}^{-1}(\boldsymbol{\eta}) + \boldsymbol{\epsilon}, \quad \mathbf{y} \sim F, \tag{1}$$

where $\mathbf{g}^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu} = \mathbb{E}(\mathbf{y}|\mathbf{X})$, with $\mathbf{g}(\cdot)$ being a suitable link function, $\boldsymbol{\epsilon}$ is an additive, unobservable error trivially defined as $\boldsymbol{\epsilon} \equiv \mathbf{y} - \mathbf{g}^{-1}(\boldsymbol{\eta})$ with $\mathbb{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}$, and F is an exponential family distribution. $\mathbf{X}$ can be written as $(\mathbf{X}^*, \mathbf{X}^+)$ where $\mathbf{X}^* = (\mathbf{X}^*_e, \mathbf{X}^*_o, \mathbf{X}^*_u)$ and $\mathbf{X}^+ = (\mathbf{X}^+_e, \mathbf{X}^+_o, \mathbf{X}^+_u)$. Symbols $*$ and $+$ indicate whether the matrix considered refers to discrete variables (such as categorical and binary variables) or continuous regressors, and subscripts e, o and u whether the matrix considered contains endogenous variables, observed confounders and unobserved confounders, respectively. The dimensions of the matrices involving the discrete regressors are $n \times p^*_e$, $n \times p^*_o$ and $n \times p^*_u$, where n denotes the sample size and the p^* the numbers of columns of the respective matrices. Following the same criterion, dimensions for the matrices containing the continuous predictors can be defined. The linear predictor of a GAM is typically given as

$$\boldsymbol{\eta} = \mathbf{X}^* \boldsymbol{\beta}^* + \sum_j \mathbf{f}_j(\mathbf{x}^+_j),$$

where $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^*_e, \boldsymbol{\beta}^*_o, \boldsymbol{\beta}^*_u)$ represents the vector of unknown regression coefficients for $\mathbf{X}^*$, and the $\mathbf{f}_j$ are unknown smooth functions of the covariates $\mathbf{x}^+_j$ in $\mathbf{X}^+$, which can be approximated using, for instance, regression splines (e.g., Marra and Radice, 2010). The generic regression spline for the jth continuous variable can be written as

$$\mathbf{f}_j(\mathbf{x}^+_j) = \mathbf{X}^+_j \boldsymbol{\theta}_j,$$

9

where $\mathbf{X}_j^+$ is the model matrix containing the spline bases (e.g., B-splines, cubic regression or thin plate regression splines) for $\mathbf{f}_j$, with parameter vector $\boldsymbol{\theta}_j$ (e.g., Wood, 2006). In order to identify (1), the $\mathbf{f}_j(\mathbf{x}_j^+)$ are subject to identifiability constraints, such as $\sum_{i=1}^{n} f_j(x_{ji}^+) = 0 \ \forall j$.

Since we can not observe $\mathbf{X}_u = (\mathbf{X}_u^*, \mathbf{X}_u^+)$, inconsistent estimates are expected. This is because the assumption that $\mathbb{E}(\mathbf{X}^{\mathsf{T}}\boldsymbol{\epsilon}) = \mathbf{0}$ would be violated. However, provided that valid IVs, $\mathbf{X}_{IV} = (\mathbf{X}_{IV}^*, \mathbf{X}_{IV}^+)$, are available, consistent estimates can be obtained by modeling the endogenous variables. Recall that valid IVs are associated with the endogenous variables conditional on the remaining covariates in the model, are independent of $\mathbf{y}$ conditional on the observed and unobserved confounders, and are independent of $\mathbf{X}_u$ (e.g., Wooldrige, 2010). In the GAM context, the endogenous variables can be modeled by using the following set of flexible auxiliary regressions (e.g., Marra and Radice, 2011a; Terza $et\ al.$, 2008)

$$\mathbf{x}_{ep} = \mathbf{g}_p^{-1}\{\mathbf{Z}_p^*\boldsymbol{\alpha}_p^* + \sum_j \mathbf{f}_j(\mathbf{z}_{jp}^+)\} + \boldsymbol{\xi}_{up}, \quad p = 1, \ldots, h, \tag{2}$$

where $\mathbf{x}_{ep}$ represents either the pth discrete or continuous endogenous predictor, $\mathbf{Z}_p^* = (\mathbf{X}_o^*, \mathbf{X}_{IVp}^*)$ with corresponding vector of unknown parameters $\boldsymbol{\alpha}_p^*$, and $\mathbf{Z}_p^+ = (\mathbf{X}_o^+, \mathbf{X}_{IVp}^+)$.

The reason why the regressions in (2) can be used to correct the parameter estimates of the equation of interest is as follows. Once the observed confounders have been accounted for and if the IVs meet the aforementioned conditions, the $\boldsymbol{\xi}_{up}$ contain information about the unobserved confounders; for a more rigorous treatment of this aspect see Marra and Radice (2011a) who, using a Taylor approximation of order 1, show that the components entering the $\boldsymbol{\xi}_{up}$ include unobserved variables. In fact, consistent estimates for the parameters in model (1) can be obtained by using the following 2SGAM approach:

1. for each endogenous variable in the model, obtain consistent estimates of $\boldsymbol{\alpha}_p^*$ and the $\mathbf{f}_j$ by fitting the corresponding equations through a GAM method. Then, calculate the quantities

$$\hat{\boldsymbol{\xi}}_{up} = \mathbf{x}_{ep} - \mathbf{g}_p^{-1}\{\mathbf{Z}_p^*\hat{\boldsymbol{\alpha}}_p^* + \sum_j \hat{\mathbf{f}}_j(\mathbf{z}_{jp}^+)\}, \quad p = 1, \ldots, h;$$

2. fit a GAM defined by

$$\mathbf{y} = \mathbf{g}^{-1}\{\mathbf{X}_{eo}^*\boldsymbol{\beta}_{eo}^* + \sum_j \mathbf{f}_j(\mathbf{x}_{jeo}^+) + \sum_p \mathbf{f}_p(\hat{\boldsymbol{\xi}}_{up})\} + \boldsymbol{\varsigma}, \tag{3}$$

where $\mathbf{X}_{eo}^* = (\mathbf{X}_e^*, \mathbf{X}_o^*)$ with parameter vector $\boldsymbol{\beta}_{eo}^*$, and $\mathbf{X}_{eo}^+ = (\mathbf{X}_e^+, \mathbf{X}_o^+)$.

In practice, the 2SGAM estimator can be implemented using GAMs represented via any penalized regression spline approach. For instance, the models in (2) and (3) can be fitted through penalized likelihood which can be maximized by penalized iteratively reweighted least squares (P-IRLS, e.g., Wood, 2006). Note that the use of roughness penalties during the model-fitting process is crucial to avoid the problem of overfitting (which is likely to occur when using a flexible model specification); this is important to produce practically useful results (e.g., Marra and Radice, 2010). Specifically, the use of the quadratic penalty $\sum_j \lambda_j \boldsymbol{\theta}^\mathsf{T} \mathbf{S}_j \boldsymbol{\theta}$, where the $\mathbf{S}_j$ are matrices measuring the roughness of the smooth functions, allows for the control of the trade-off between fit and smoothness through the smoothing parameters λ_j, which can be estimated as explained in Wood (2006).

The presence of a relationship between the response and unobservables that are associated with the endogenous predictors can lead to inconsistent estimates of the impacts of the latter variables. But because the $\mathbf{f}_p(\hat{\boldsymbol{\xi}}_{up})$ in (3) allows us to flexibly account for the impacts of unobserved confounders on the response, the linear/non-linear effects of the endogenous regressors can be estimated consistently. Note that the $\hat{\mathbf{f}}_p(\hat{\boldsymbol{\xi}}_{up})$ will contain a mixture of effects which makes them not interpretable. However, this is not problematic since all that is required is to account for the presence of unobservables. See Marra and Radice (2011a) for full details.

Confidence intervals for the components of a GAM can be constructed using the well known Bayesian 'confidence' intervals originally proposed by Wahba (1983) in the univariate spline model context, and then generalized to the component-wise case when dealing with GAMs (e.g., Gu, 2002; Marra and Wood, 2012; Wood, 2006). The reason why such Bayesian intervals are referred to as 'confidence' intervals is because they include both a bias and variance component, a fact which makes them have good observed *frequentist* coverage probabilities across the function (Marra and Wood, 2012). For simplicity of exposition and without loss of generality, let us consider a generic GAM whose linear predictor is made up of smooth terms only. The large sample posterior for the parameter vector containing all spline coefficients is $\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda}, \phi \sim \mathcal{N}(\hat{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is

the maximum penalized likelihood estimate of $\boldsymbol{\theta}$, which is of the form $(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{z}$, $\mathbf{V}_{\boldsymbol{\theta}} = (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X} + \mathbf{S})^{-1}\phi$, $\mathbf{X}$ contains the columns associated with the regression spline bases for the $\mathbf{f}_j$, $\mathbf{W}$ and $\mathbf{z}$ are the diagonal weight matrix and the pseudodata vector at convergence of the P-IRLS algorithm and $\mathbf{S} = \sum_j \lambda_j \mathbf{S}_j$. Now, because the second-stage of 2SGAM can not account for the additional source of variability introduced via the residuals calculated in the first step, the intervals for the components in the second-step model will be too narrow, hence leading to poor coverage probabilities. This can be rectified via posterior simulation (Marra and Radice, 2011a). Intuitively, samples from the posterior distribution of each first-step model are used to obtain samples from the posterior of the quantities of interest $\boldsymbol{\xi}_{up}$. Then, given N_b vector replicates for each $\boldsymbol{\xi}_{up}$, N_d random draws from the N_b posterior distributions of the second-stage model are used to construct approximate intervals for the smooth functions. Marra and Radice (2011a) suggest that $N_b = 25 \times p$ and $N_d = 100$ yield good coverage probabilities. These settings will be used to construct 95% intervals for the function estimates obtained using 2SGAM.

3.2   Applying 2SGAM to the fertility data

The regression model for `children` can be written as

$$\texttt{children} \sim \text{Poisson}(\boldsymbol{\mu}),$$

$$\texttt{children} = \underbrace{\exp\left\{\mathbf{X}_o^*\boldsymbol{\beta}_o^* + f_{\texttt{educ}}(\texttt{education}) + f_{\texttt{age}}(\texttt{age}) + \text{unobservables}\right\}}_{\boldsymbol{\mu}} + \boldsymbol{\epsilon},$$

where, to make the link with the general framework presented in the previous section, the chosen link function $\mathbf{g}$ is log, $\mathbf{X}_e^+ = \texttt{education}$, $\mathbf{X}_e^*$ is an empty matrix (as there is only one continuous endogenous variable), $\mathbf{X}_o^*$ contains all binary and categorical variables listed in Table 1 and $\mathbf{f}_j(\mathbf{x}_j^+) = f_{\texttt{age}}(\texttt{age})$. Note that empirical models that investigate the determinants of fertility in Malawi have usually included education as a categorical variable (Kalipeni, 1997; Kazembe, 2009). If we were to impose *a priori* assumptions by categorizing years of schooling, we might fail to capture some features in the fertility-education relationship. For instance, the choice of cut-off points for the categorization of this variable is typically determined by the researcher's choices rather than by objective factors. In contrast, the flexible modeling of education as a continuous

covariate allows us to capture all of the informative power of the variable without making any *a priori* assumptions about the linearity or non-linearity of the relationship.

The naive model is represented by the following generalised additive Poisson regression

$$\texttt{children} = \exp\left\{\mathbf{X}_o^* \boldsymbol{\gamma}_o^* + f_{\texttt{educ}}(\texttt{education}) + f_{\texttt{age}}(\texttt{age})\right\} + \boldsymbol{\psi}. \tag{4}$$

The first-stage regression of the 2SGAM approach is the additive model

$$\texttt{education} = \texttt{firsthalf}\alpha_{IV} + \mathbf{X}_o^* \boldsymbol{\alpha}_o^* + \boldsymbol{\xi}_u + f_{1,\texttt{age}}(\texttt{age}) + \boldsymbol{\xi}, \tag{5}$$

where $p = 1$ (because there is only one endogenous variable), $\mathbf{X}_{IV}^* = \texttt{firsthalf}$ and $\mathbf{X}_{IV}^+$ is empty. The second-stage model is a generalized Poisson additive regression

$$\texttt{children} = \exp\left\{\mathbf{X}_o^* \boldsymbol{\beta}_o^* + f_{2,\texttt{educ}}(\texttt{education}) + f_{2,\texttt{age}}(\texttt{age}) + f_\xi(\hat{\boldsymbol{\xi}})\right\} + \boldsymbol{\varsigma}. \tag{6}$$

In all of the models, the smooth functions of the continuous covariates are based on thin plate regression splines with a second-order penalty and ten regression bases (Wood, 2006), and all binary and categorical variables are entered into the models as parametric predictors.

The chosen IV is a binary variable that indicates whether a woman was born in the first half of the year (`firsthalf`). This variable may be considered a valid instrument on the following grounds. 1) This variable is logically and statistically associated with education. Women born during the first semester of the year are expected to start school later and hence have a lower educational attainment level than those born during the rest of the year. Some descriptive statistics by semester and year of birth of women (available upon request) empirically demonstrate that the average number of years of education is generally higher for women born during the second half of the year than for those born during the first half of the year. 2) It is reasonable to assume that `firsthalf` does not have an effect on fertility, given the confounders in the model. 3) It is unlikely that `firsthalf` is associated with unobservable confounders such as ability, motivation and preference. This instrument has been used in similar contexts, such as in studies estimating the education-fertility relationship in Botswana (e.g., Sobotka *et al.*, 2013; Marra and Radice, 2011b; Wooldridge, 2010). Of course, because some of the IV assumptions are not testable, it is

impossible to be entirely sure of the validity of the instrument. One potential criticism of this IV choice might be that it is weak (Bound *et al.*, 1995) and could lead to inconsistent estimates. However, this point will be addressed in the next section. The 2SGAM approach described in this section has been implemented in the `R` environment.

4   Results

We begin the discussion of the empirical results by comparing the naive estimates with those of the 2SGAM. The estimated effects of `age` and `education` on Malawian women's number of children are displayed in Figure 2, and the estimates of the binary and categorical covariates are shown in Table 2. The results obtained using the naive approach show that the education-fertility relationship follows a non-linear decreasing trend that is particularly strong for women with more than eight years of education. This pattern confirms the findings of Kalipeni (1997) and Kazembe (2009). However, one limitation of these studies is that either education was integrated into the models as a categorical variable or a linear relationship with the outcome variable was assumed, hence masking potentially important non-linearities. Similar evidence of the presence of an inverse relationship between fertility and education has been found in other African countries, including Botswana, Morocco, Senegal, Sierra Leone, Uganda, Zambia, and Zimbabwe (e.g., Martin, 1995; Lutz and Kc, 2011). Lutz and Kc (2011) noted that this pattern is typically more pronounced in countries that have high fertility levels and that are in the early stages of demographic transition. In particular, some scholars suggest that more educated women have greater autonomy in reproductive decision-making and engage in more intensive parenting to ensure a better quality of life for their children (e.g., Martin, 1995; Kravdal, 2000; Basu, 2002; Manda and Meyer, 2005; Stephenson *et al.*, 2007; Lutz and Kc, 2011).

The 2SGAM estimated curve for education differs from that of the naive model. More specifically, the 2SGAM results suggest the existence of an inverted U-shaped relationship between fertility and education with a turning point around eight years of schooling. The estimates indicate that the first eight years of schooling may increase women's number of children. Conversely, the number of years of education that a woman completes beyond this threshold reduces fertility (see the second plot of Figure 2). We also observe that for women who have up to five years

of schooling, the naive model overestimates the effect of education on children more than the 2SGAM model does. However, the estimated effects of the inverted U-shaped relationship show that although the first nine years of schooling have a statistically significant effect on the outcome variable, beyond this threshold, statistical significance is weak or absent (also see Table 4 in the Appendix). In light of this issue, additional assumptions will be empirically investigated in the next section.

Evidence of an inverted U-shaped relationship has already been found for the African countries of Burundi, Kenya, and Liberia, where descriptive statistics were employed (e.g., Martin, 1995). For Malawi, however, the naive model did not yield an inverted U-shaped curve, and 2SGAM was used to identify this possible pattern. In the literature, scholars have attributed this atypical pattern to several factors. For example, some studies suggest that in the absence of conscious birth control, a modest improvement in education may increase fertility in women as a result of reduced breastfeeding and postpartum abstinence (e.g., Martin, 1995). Lower fertility is likely among better-educated women (i.e., those who complete additional years of education above the threshold level of eight to ten years of schooling) because their knowledge and skills allow them to perform jobs in sectors that are incompatible with child-rearing (e.g., Birdsall and Griffin, 1988; Kalipeni, 1997). Therefore, it is reasonable to conclude that better-educated women have greater motivation to use birth control than do less educated women given the higher opportunity costs of unplanned childbearing.

Following Sobotka *et al.* (2013), we tested for the presence of endogeneity via the hypothesis $H_0 : f_\xi(\hat{\boldsymbol{\xi}}) = \mathbf{0}$, using estimates from second-stage (6). We found a p-value of 1.13^{-06}. This value provides evidence to reject the null hypothesis of no endogeneity. Therefore, a more complex modeling approach is justified over simpler model (4). To obtain consistent estimates, it is important that the instrument is strongly associated with the endogenous variable. Toward this end, we employed the Cragg and Donald (1993) statistic to test the hypothesis that `firsthalf` is a strong instrument. Because there is only one endogenous variable, the Cragg-Donald test is the F-statistic in the first-stage regression (5). The statistic value is 192, which is well beyond the value of 10 that Staiger and Stock (1997) suggest as a threshold for rejecting the null hypothesis (see also Table 3 in the Appendix). These testing tools suggest that the education-fertility relationship is affected by the presence of a mix of unobserved confounding effects, such as abilities,

motivations, and preferences, which are not accounted for in the naive model.

Regarding the observed confounders, the first plot in Figure 2 suggests that the women's number of children increases in a non-linear manner throughout their reproductive lives (between 15 and 49 years of age). More specifically, the degree of increase is stronger for women between the ages of 15 and 25 and then progressively stabilises until the end of women's reproductive lives (i.e., between 45 and 50 years). The same result is obtained using both estimation procedures.

INSERT FIGURE 2 HERE

The estimates of the remaining observed confounding effects, which are reported in Table 2, highlight that `religion` and `never_married` have greater effects in the 2SGAM than they do in the naive model and that the coefficient for `region` has differing signs for the two models. It is worth noting that although the variable of interest is education, the effects of the control variables on children should also change (to some degree) when 2SGAM is used. This is not surprising because the error term of the `children` equation and `education` are not independent, hence inducing bias in all parameters of the model (e.g., Wooldrige, 2010, ch. 5).

INSERT TABLE 2 HERE

4.1   The joint effect of woman's age and education on fertility

The analysis presented in the previous section does not allow us to analyse the effects of `education` on `children` according to the women's birth cohorts. This issue arises because the sample used for our study includes women who are at the beginning of their childbearing years as well as women who are concluding their childbearing period. To address this matter, we estimate the joint effects of `age` and `education` on `children`.

More specifically, naive model (4) becomes

$$\texttt{children} = \exp\left\{\mathbf{X}_o^* \boldsymbol{\gamma}_o^* + f_{\texttt{educ}}(\texttt{education}) + f_{\texttt{age}}(\texttt{age}) + f_{\texttt{educ},\texttt{age}}(\texttt{education},\texttt{age})\right\} + \psi. \quad (7)$$

Using the 2SGAM approach, second-stage model (6) becomes

$$\texttt{children} = \exp\left\{\mathbf{X}_o^*\boldsymbol{\beta}_o^* + f_{2,\texttt{educ}}(\texttt{education}) + f_{2,\texttt{age}}(\texttt{age}) + f_{2,\texttt{educ},\texttt{age}}(\texttt{education}, \texttt{age}) + f_\xi(\hat{\boldsymbol{\xi}})\right\} + \varsigma,$$

(8)

where $f_{\texttt{educ},\texttt{age}}(\texttt{education}, \texttt{age})$ estimates the joint effects of $\texttt{age}$ and $\texttt{education}$ on $\texttt{children}$. For the methodological details regarding the construction of a smooth component of two covariates, we refer the reader to Wood (2006).

INSERT FIGURE 3 HERE

INSERT FIGURE 4 HERE

INSERT FIGURE 5 HERE

Here, we only present the joint effects of $\texttt{age}$ and $\texttt{education}$ on $\texttt{children}$, which are our primary focus. The estimated effects of the observed confounding variables and the results of the test for endogeneity lead to similar conclusions as those discussed in the previous section and are available upon request. We should note that estimating model (8) confirms that endogeneity is a concern in the analysis of the fertility-education relationship in Malawi. The main findings obtained using 2SGAM are reported in Figure 3 and show that education has an inverted U-shaped relationship to fertility that is particularly marked among women who are more than 30 years old. This pattern is not captured in the estimation of the fertility-education relationship that is made using naive model (7). When endogeneity is not taken into account, an inverse relationship is detected (see the first two graphs of Figure 3).

To deepen our analysis, we also investigate the hypothesis that the effect of $\texttt{education}$ on $\texttt{children}$ may vary by women's birth cohort and area of residence (i.e., rural or urban). We examine this assumption by estimating models (7) and (8) for the two separate samples of women living in rural and urban areas. The main results are reported in Figures 4 and 5.

17

Focusing on the results of 2SGAM, we observe that endogeneity is a concern that affects the fertility-education relationship for women living in rural areas but does not affect that relationship for those who live in urban areas (the results of the endogeneity test are available upon request). More specifically, an accentuated, inverted U-shaped fertility-education relationship with a turning point around eight years of schooling is observed among women living in a rural setting who are more than 30 years old. This same relationship is substantially flatter for younger women. Furthermore, we observe the existence of an inverse relationship characterised by non-linearities that vary by birth cohort for women living in an urban context. In general, these findings highlight the complexity of the fertility-education relationship in Malawi.

5 Conclusion

High fertility coupled with declining death rates is driving a rapid increase in the population in Malawi, where the average annual population growth exceeds 2%. This demographic trend is a critical socio-economic issue for Malawi because it coincides with inadequate resources and a lack of efficient policies oriented towards the country's development. Therefore, slower population growth and a sustainable level of fertility should be pursued through policy initiatives.

This study analysed the relationship between education and fertility among Malawian women, accounting for both observed and unobserved confounders. This analysis is motivated by observational studies that have identified education as a key contributor to the decline of fertility. Our investigation is based on micro data from the 2010 MDHS, which contains information on the number of children per woman. The study takes into account education and other socio-demographic characteristics for a sample of women between the ages of 15 and 49 years. One initial finding of our study is that women's education has a significant impact on their number of children. Another important finding is that endogeneity is likely to be a concern in research on this relationship. The naive estimates for the entire sample of women indicated that education has a negative and non-linear effect on the number of children. However, when endogeneity is accounted for using the 2SGAM approach, the impact of education on the outcome variable exhibits an inverted-U shape. We also observe that in this pattern, the first nine years of schooling have a statistically significant impact on the outcome variable, whereas beyond this threshold, the confidence intervals become

18

progressively larger and indicate that the relationship has weak (or no) statistical significance. This evidence has motivated us to conduct further empirical analysis. As a result, we now know that the effect of education on the number of children that a woman has varies based on her birth cohort and area of residence (whether rural or urban). The relationship appears to be affected by endogeneity for women living in rural areas but not for those living in urban areas. Notably, our analysis suggests that unobservable confounding factors such as abilities, motivations, and preferences contribute to the formation of an inverted U-shaped relationship between education and fertility for women who are more than 30 years old and who live in rural areas, whereas for women living in urban areas in which endogeneity is not an issue, the fertility-education relationship exhibits a non-linear decreasing trend.

As suggested in the literature, one possible interpretation of the atypical inverted U-shaped relationship is that in the absence of conscious birth control, fewer years of education have the potential to increase fertility as a result of reduced breastfeeding and postpartum abstinence. A reduction in fertility is found for women who achieve the highest levels of schooling because, for example, better education helps women to acquire greater knowledge and the skills that they require to find jobs in sectors that are incompatible with child-rearing. In light of this explanation, it is reasonable to believe that better-educated women are more likely to use birth control methods (e.g., modern contraceptive methods) than are women with lower levels of education (Zanin *et al.*, 2014). However, we are not able to determine whether the inverted-U shape that is observed for women older than 30 who live in rural areas is generalisable to younger generations who advance. In general, our findings suggest that in Malawi, the environment in which a woman lives (whether rural or urban) is an important factor in the relationship between fertility and education in her life. Based on these findings, policy-makers should orient their action strategies towards programmes that contribute to increased access to, quality of, and achievement of education for a larger number of women[1] (see also Fan and Zhang, 2013), as well as promote educational initiatives on birth control, especially for women living in rural areas. Obviously, reducing fertility through improvements in women's education is a strategy that will require time and investment. As next steps, researchers should pursue a better understanding of the complex relationship between

---

[1]As a non-exhaustive list of policies, we can cite (i) compulsory primary education, (ii) investment and improvements in the quality of the school infrastructure, (iii) efforts to balance the teacher-student ratio, and (iv) support for the purchase of school supplies for students and teachers (e.g., textbooks, pens, pencils).

fertility and education using longitudinal data, which were not available for our analysis. Moreover, future studies should consider collecting information on the quality of schooling, family well-being, and women's lifestyle and use of time. These variables are important because considering them may help us to better interpret the investigated relationship and to develop specific related policies.

Appendix

For the sake of completeness, below we report the estimation results that indicate the parametric and non-linear effects for first-stage model (5).

INSERT TABLE 3 HERE

INSERT FIGURE 6 HERE

INSERT TABLE 4 HERE

References

[1] Angeles G, Guilkey DK, Mroz TA (2005) The effects of education and family planning programs on fertility in Indonesia, Econ Dev Cult Change 54:165–202

[2] Angrist JD, Krueger AB (1991) Does compulsory school attendance affect schooling and earning?, Q J ECON 106:979–1014

[3] Barrera A (1991) The interactive effects of mother's schooling and unsupplemeted breastfeeding on child health, J Dev Econ 34:81–98

[4] Bailey (1989) Female education and fertility in rural Sierra Leone: a test of the threshold hypothesis, Can Stud Popul 16:87–112

[5] Basu MA (2002) Why does education lead to lower fertility? A critical review of some of the possibilities, World Dev 30:1779–1790

[6] Blanc A, Rutenberg N (1990) An assessment of the quality of data on age at first intercourse, age at first marriage, and age at first birth in the Demographic and Health Surveys, in Assessment of DHS-I Data Quality. Columbia, Md: Institute for Resource Development/Macro System 3979

[7] Bhalotra S, Rawlings SB (2011) Intergenerational persistence in health in developing countries: the penalty of gender inequality?, J Public Econ 95:286–299

[8] Birdsall MN, Griffin CC (1988) Fertility and poverty in developing countries, J Policy Model 10:29–55

[9] Bound J, Jaeger DA, Baker RM (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, J Am Stat Assoc 90:443–450

[10] Chen Y, Li H (2009) Mother's education and child health: is there a nurturing effect?, J Health Econ 28:413–1426

[11] Chimombo JPG (2005) Quantity versus quality in education: case studies in Malawi, Int Rev Educ 51:155–172

[12] Colclough C (1982) The impact of primary schooling on economic development: a review of the evidence, World Dev 10:167–185

[13] Cohen B (1998) The emerging fertility transition in Sub-Saharan Africa, World Dev 26:1431–1461

[14] Cragg JG, Donald SG (1993) Testing identifiability and specification in instrumental variable models, Economet Theor, 9:222–240

[15] Fan SC, Zhang J (2013) Differential fertility and intergenerational mobility under private versus public education, J Popul Econ, 26:907–941

[16] Gu C (2002) *Smoothing spline ANOVA models*. London: Springer-Verlag.

[17] Hausman J (1978) Specification tests in econometrics. Econometrica 46:1251–1271

[18] Harwood-Lejeune AL (2003) Rising age at marriage and fertility in Southern and Easter Africa, Eur J Popul 17:261–280

[19] Heaton TB (2011) Does religion influence fertility in developing countries, Popul Res Policy Rev 30:449–465

[20] IMF [International Monetary Found] (2012) World Economic Outlook Database for April 2012. http://www.imf.org/external/ns/cs.aspx?id=28. Accessed 1th Mai 2012

[21] ICF International (2011) Demographic and Health Surveys Methodology - Questionnaires: Household, Woman's, and Man's. MEASURE DHS Phase III (project, implemented from 2008-2013): Calverton, Maryland, USA. http://www.measuredhs.com/publications/publication-DHSQ6-DHS-Questionnaires-and-Manuals.cfm. Accessed 1th Mai 2012

[22] Kadzamira E, Rose P (2003) Can free primary education meet the needs of the poor? Evidence from Malawi, Int J Educ Dev 23:501–516

[23] Kalipeni E (1997) Population pressure, social change, culture and Malawi's pattern of fertility transition, Afr Stud Rev 40:173–208

[24] Kazembe LN (2009) Modelling individual fertility levels in malawian women: a spatial semi-parametric regression model, Stat Method Appl 18:237–255

[25] Kravdal O (2000) A search for aggregate-level effects of education on fertility, using data from Zimbabwe, Demogr Res 3:1–34

[26] Lutz W, Kc S (2011) Global human capital: integrating education and population, Science 333:587–592

[27] MDHS [Malawi Demographic and Health Survey] (2012) Country Quickstats. http://www.measuredhs.com/Where-We-Work/Country-Main.cfm?ctry_id=24&c=Malawi&Country=Malawi&cn=. Accessed 1th Mai 2012

[28] MMDG [Malawi Millennium Development Goals] (2010) Malawi Millennium Development Goals. http://planipolis.iiep.unesco.org/upload/Malawi/MalawiMDGs2010Report.pdf

[29] Manda S, Meyer R (2005) Age at first marriage in Malawi: a bayesian multilevel analysis using discrete time-to-event model, J R Stat Soc Ser A 168:439–455

[30] Martin TC (1995) Women's education and fertility: results from 26 demographic and health survey, Stud Family Plann 26:187–202

[31] Marra G, Radice R (2010) Penalised regression splines: theory and application to medical research, Stat Methods Med Res 19:107–125

[32] Marra G, Radice R (2011a) A flexible instrumental variable approach, Stat Model 11:581–603

[33] Marra G, Radice R (2011b) Estimation of a semiparametric recursive bivariate probit model in the presence of endogeneity, Can J Stat 39:259–6279

[34] Marra G, Wood S (2011) Pratical variable selection for generalized additive models, Comput Stat Data An 55:2372–2387

[35] Marra G, Wood S (2012) Coverage Properties of Confidence Intervals for Generalized Additive Model Components, Scand J Stat 39:53–74

[36] Moyi P (2010) Household characteristics and delayed school enrollment in Malawi, Int J Educ Dev 30:236–242

[37] National Statistical Office and ICF Macro [NSO and ICF macro] (2011) Malawi Demographic and Health Survey 2010. Zomba, Malawi, and Calverton, Maryland, USA: NSO and ICF Macro

[38] Nychka D (1988) Bayesian confidence intervals for smoothing splines, J Am Stat Assoc 83:1134–1143

[39] Osili UO, Long BT (2008) Does female schooling reduce fertility? Evidence from Nigeria, J Dev Econ 87:57–75

[40] Palamuleni ME (2011) Socioeconomic determinants of age at marriage in Malawi, Int J Sociol Anthropol 3:224–235

[41] Schafer MJ (2006) Household change and rural school enrollment in Malawi and Kenya, Sociol Quart 47:665–691

[42] Sobotka F, Radice R, Marra G, Kneib T (2013), Estimating the relationship of women's education and fertility in Botswana using an instrument variable approach to semiparametric expectile regression, J R Stat Soc Ser C 62:25–45

[43] Staiger D, Stock JH (1997) Instrumental variables regression with weak instruments, Econometrica 65:557-586

[44] Stephenson R, Baschieri A, Clements S, Hennink M, Madise N (2007) Contextual influences on modern contraceptive use in Sub-Saharan Africa, Am J Public Health 97:1233–1240

[45] Strauss J, Thomas D (1995) *Human resources: Empirical modeling of household and family decisions*, in J. Behrman and T.N. Srinivasan, eds. The Handbook of Development Economics, 3A. Amsterdam: Elsevier

[46] Strulik H, Vollmer S (2013) The fertility transition around the world, J Popul Econ. doi: 10.1007/s00148-013-0496-2

[47] Terza JV, Basu A, Rathouz PJ (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling, J Health Econ 27:531-543

[48] Wahba G (1983) Bayesian 'confidence intervals' for the cross-validated smoothing spline, J R Stat Soc Ser B 45:133–150

[49] Wood S (2006) *Generalized additive models: an introduction with R.* London: Chapman & Hall

[50] Wooldrige JM (2010) *Econometric analysis of cross section and panel data.* Cambridge: MIT Press

[51] World Bank (2012) Indicators for Malawi. http://data.worldbank.org/country/malawi. Accessed 3th Mai 2012

[52] Zanin L, Marra G (2012) A comparative study of the use of generalized additive model and generalized linear models in tourism research, Int J Tour Res 14:451–468

[53] Zanin L, Radice R, Marra G (2014) A comparison of approaches for estimating the effect of women's education on the probability of using modern contraceptive methods in Malawi, SOC SCI J. In press
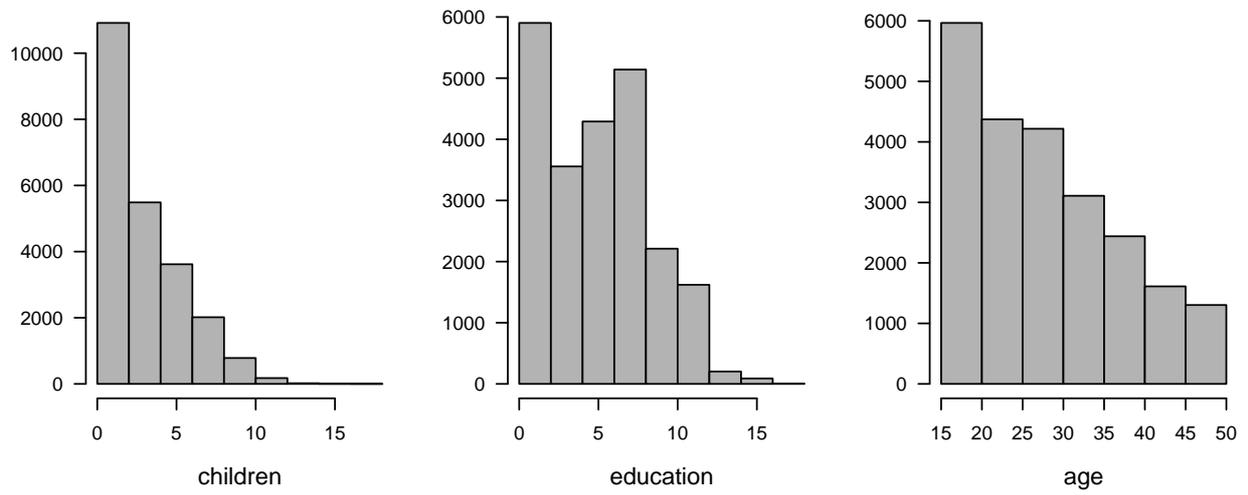
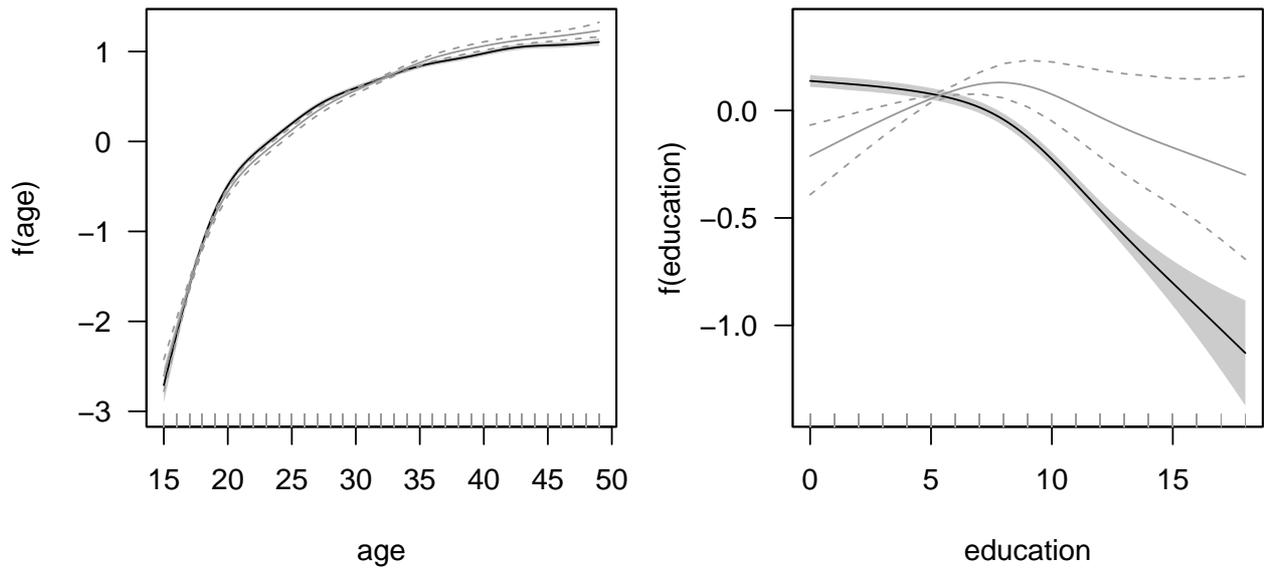Figure 1: Frequency distributions of some of the variables considered in the fertility-education study

Figure 2: Estimated non-linear effects on the scale of the linear predictor for the continuous covariates `age` and `education` included in models (4) and (6). Naive curve estimates with corresponding 95% intervals are represented by black lines and shaded regions, respectively. The results obtained using 2SGAM are reported in gray with corrected 95% intervals represented by dashed lines. The rug plot, at the bottom of each graph, shows the covariate values.
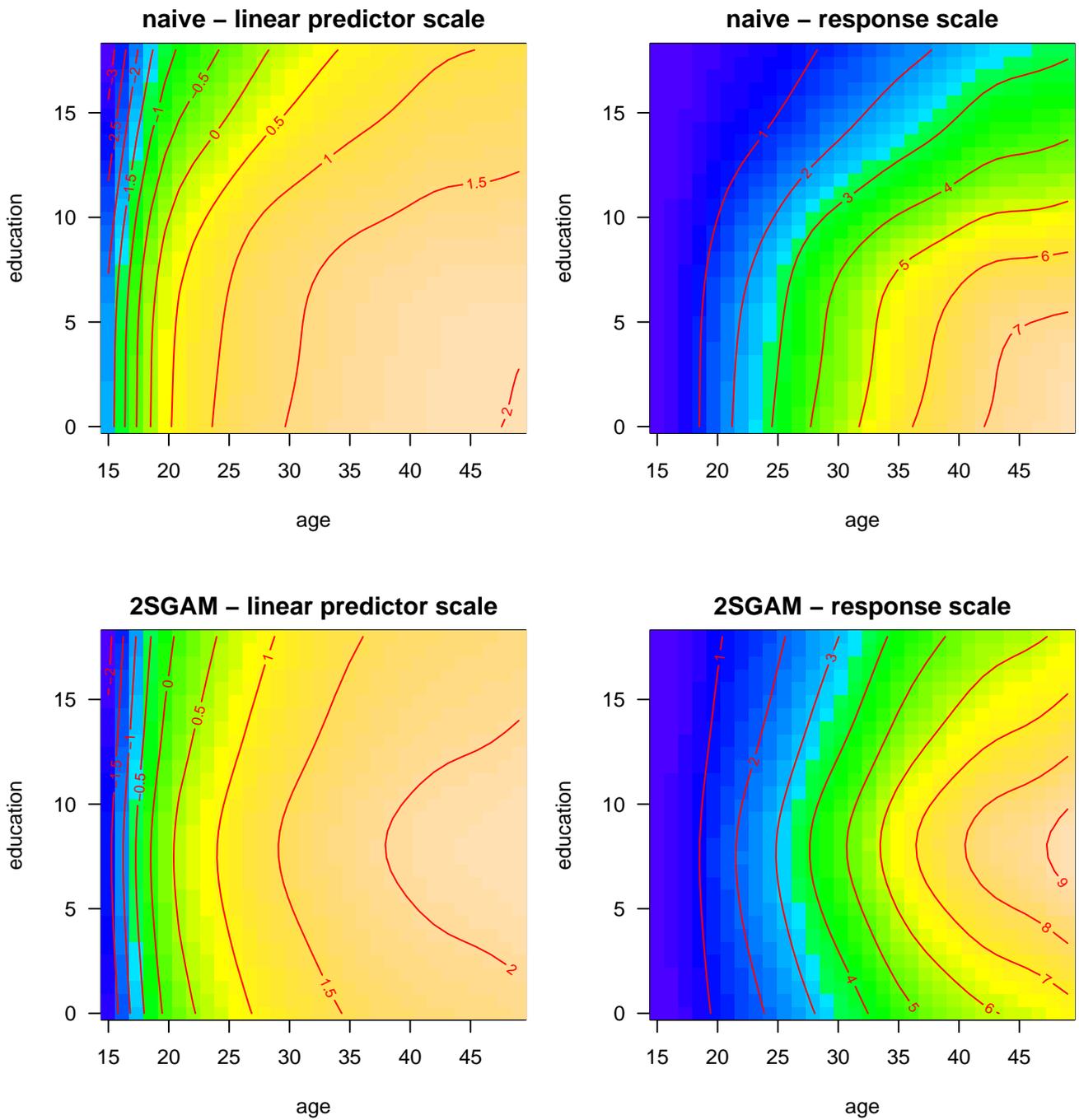
Figure 3: Contour plots of the joint impact of `age` and `education` on `children` when the naive (7) and 2SGAM (8) models are used to analyse the whole sample of women. The p-values associated with the estimated smooth functions are all smaller than 0.05, hence supporting the hypothesis of statistical significance at the 5% level. The estimated effects of the observed covariates included in the models are available upon request.

Figure 4: Contour plots of the joint impact of `age` and `education` on `children` when the naive (7) and 2SGAM (8) models are used to analyse the sample of women living in rural areas. The p-values associated with the estimated smooth functions are all smaller than 0.05, hence supporting the hypothesis of statistical significance at the 5% level. The estimated effects of the observed covariates included in the models are available upon request.
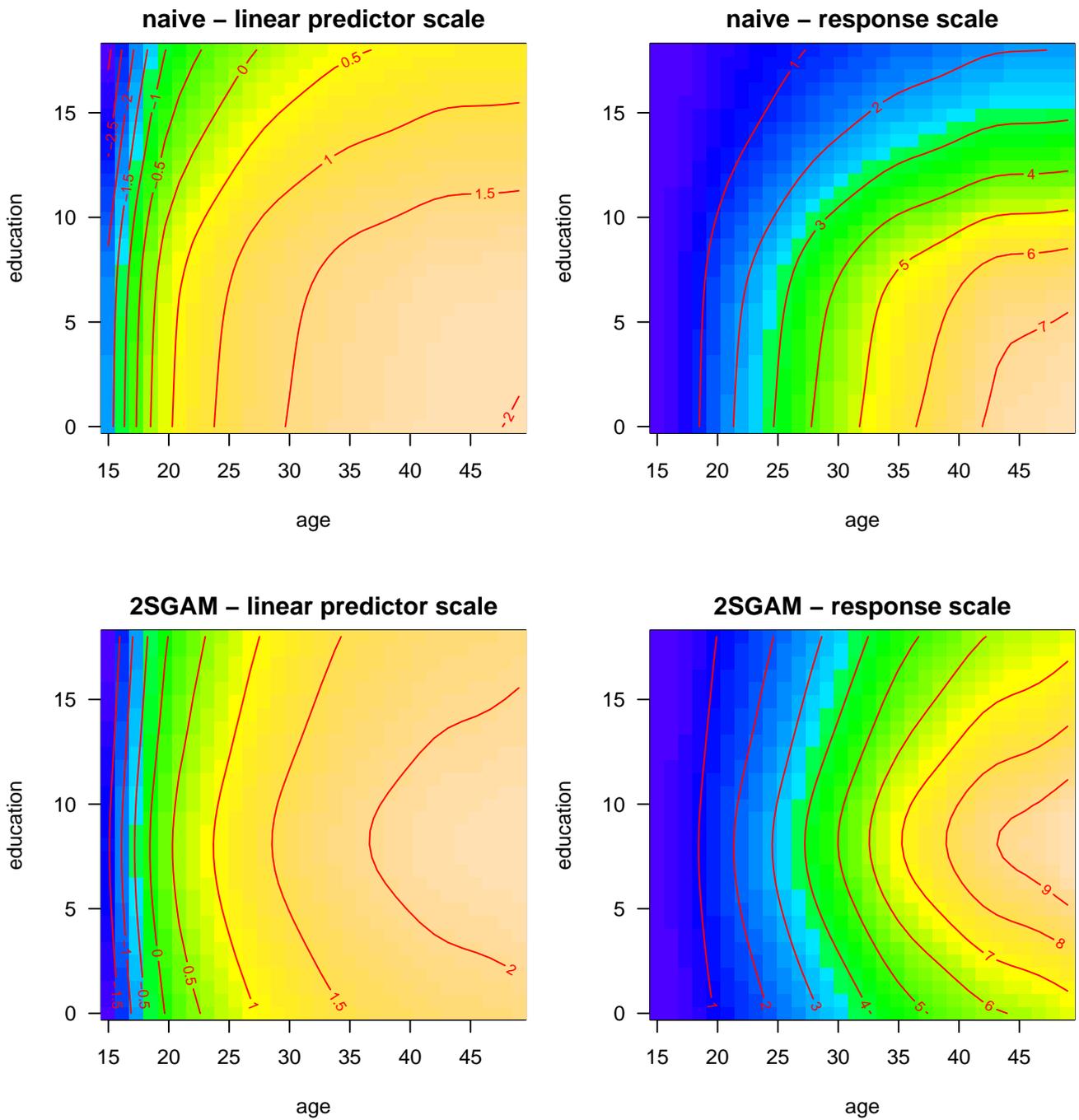
Figure 5: Contour plots of the joint impact of `age` and `education` on `children` when the naive (7) and 2SGAM (8) models for the sample of women living in urban areas. The p-values associated with the estimated smooth functions are all smaller than 0.05, hence supporting the hypothesis of statistical significance at the 5% level. The estimated effects of the observed covariates included in the models are available upon request.
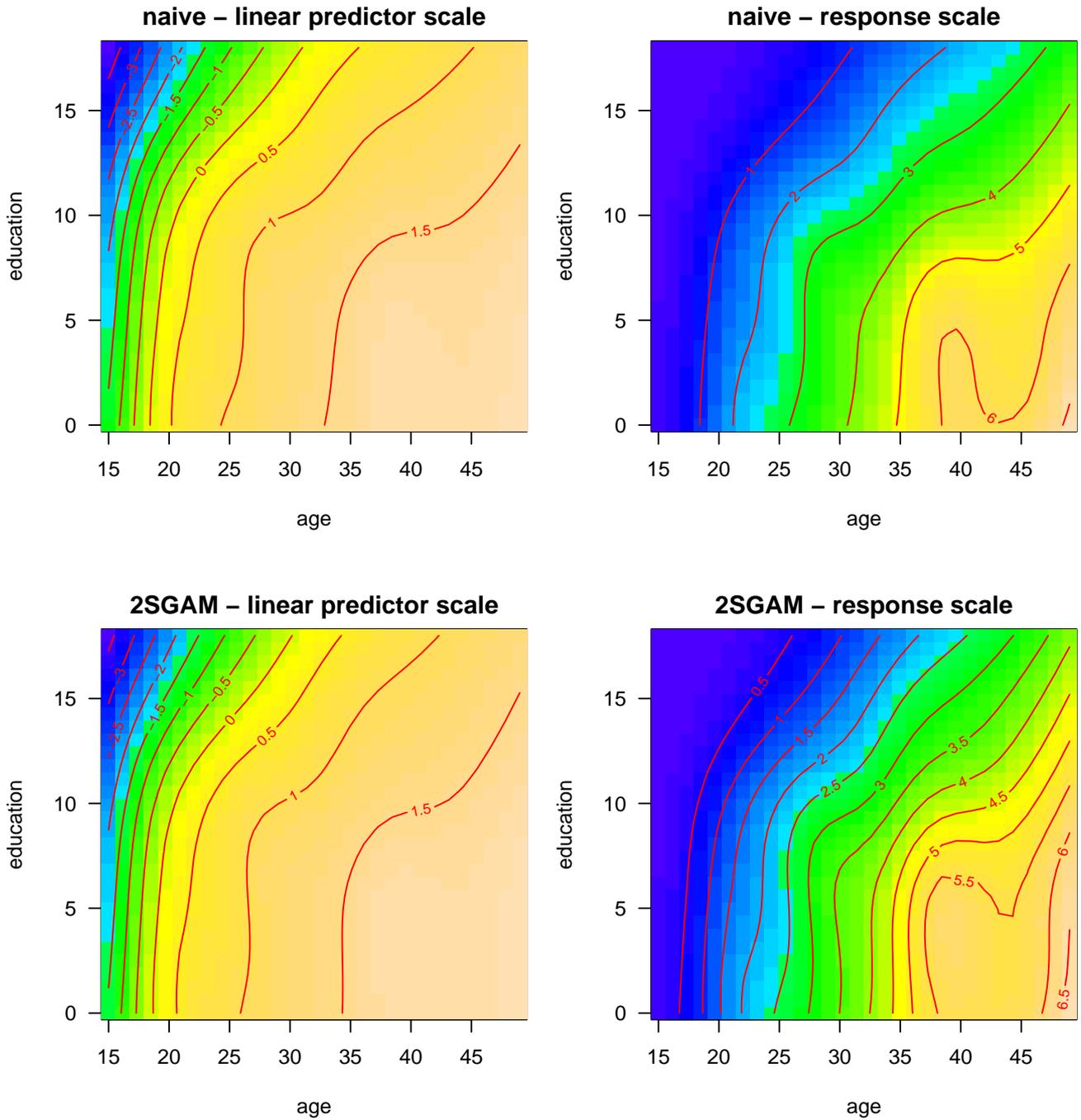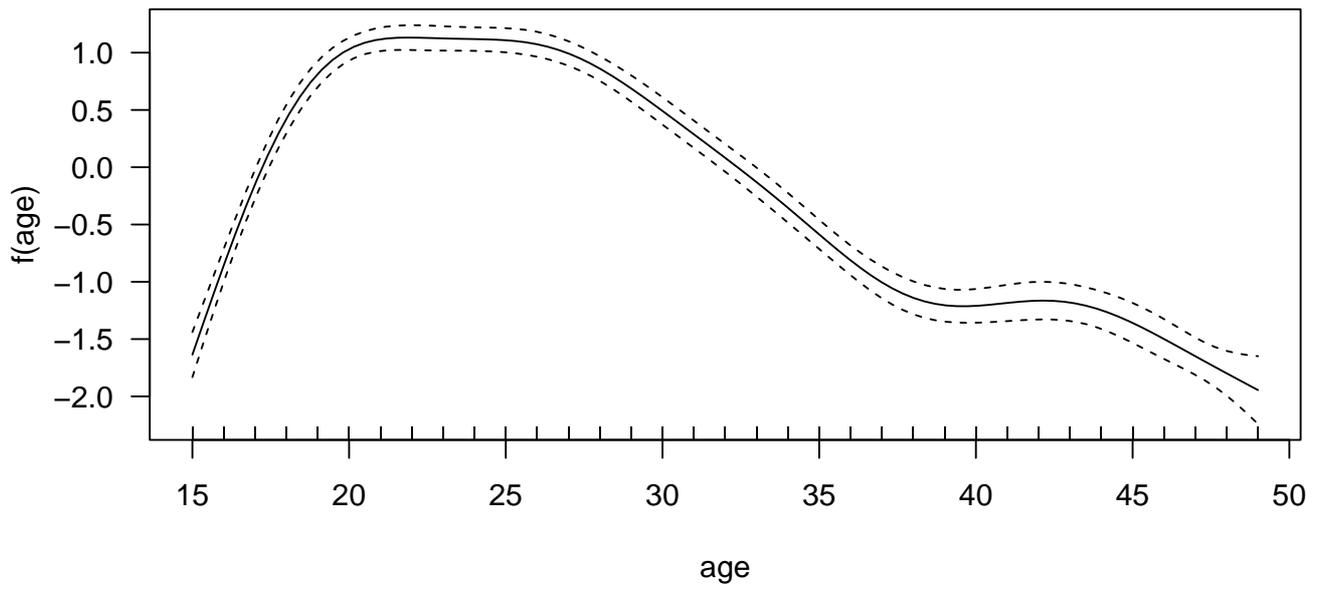
Figure 6: Estimated non-linear effects on the scale of the linear predictor for the continuous covariate `age` included in model (5). 95% confidence intervals are represented by dashed lines. The rug plot, at the bottom of each graph, shows the covariate values.

Table 1: Summary statistics of the binary and categorical variables considered in the fertility-education study

Variable	Frequency	Percentage
Region		
Northern	4189	18.20
Central	7862	34.15
Southern	10969	47.65
Urban		
No	19952	86.67
Yes	3068	13.33
Religion		
Muslim	2530	10.99
Catholic	4670	20.29
Protestant	15614	67.83
Other	206	0.89
Never married		
No	18494	80.34
Yes	4526	19.66
Observations	23020	100.00

Table 2: Estimated parametric effects for the naive and 2SGAM methods

Variable	Naive		2SGAM	
	Coefficients eq. (4)	CI	Coefficients eq. (6)	CI
Intercept	0.828	(0.792;0.864)	0.815	(0.768;0.848)
Region				
Northern	-	-	-	-
Central	-0.025	(-0.047;-0.002)	0.106	(0.051;0.180)
Southern	-0.064	(-0.086;-0.042)	0.059	(-0.006;0.128)
Urban				
No	-	-	-	
Yes	-0.095	(-0.121;-0.070)	-0.271	(-0.359;-0.192)
Religion				
Muslim	-	-	-	
Catholic	-0.027	(-0.055;0.000)	-0.125	(-0.175;-0.072)
Protestant	-0.024	(-0.048;-0.001)	-0.085	(-0.112;-0.044)
Other	-0.038	(-0.1109;0.033)	-0.009	(-0.056; 0.104)
Never Married				
No	-	-	-	-
Yes	-1.693	(-1.793;-1.592)	-1.853	(-1.999;-1.741)

CIs are 95% confidence intervals. Sample size is 23020.

Table 3: Estimated parametric effects for model (5)

Variable	2SGAM Coefficients eq. (5)	CI
Intercept	5.629	(5.460;5.798)
Region		
Northern	-	-
Central	-2.086	(-2.203;-1.970)
Southern	-1.945	(-2.057;-1.833)
Urban		
No	-	
Yes	2.691	(2.572;2.809)
Religion		
Muslim	-	
Catholic	1.395	(1.244;1.547)
Protestant	0.783	(0.650;0.915)
Other	-0.961	(-1.401;-0.521)
Never Married		
No	-	-
Yes	2.708	(2.548;2.867)
IV (firsthalf)	-0.571	(-0.652;-0.490)

CIs are 95% confidence intervals. Sample size is 23020.

Table 4: Estimated non-linear effects on the scale of the linear predictor for the continuous covariate education when the naive (4) and 2SGAM (6) models are used

Education	Naive	2SGAM
0	0.138 (0.122;0.153)	-0.212 (-0.53;-0.071)
1	0.129 (0.114;0.143)	-0.155 (-0.270;-0.040)
2	0.120 (0.104;0.136)	-0.099 (-0.188;-0.010)
3	0.109 (0.095;0.123)	-0.044 (-0.107;0.0019)
4	0.095 (0.082;0.109)	0.008 (-0.030;0.045)
5	0.077 (0.064;0.090)	0.055 (0.039;0.071)
6	0.052 (0.039;0.065)	0.095 (0.073;0.116)
7	0.015 (0.002;0.027)	0.122 (0.077;0.167)
8	-0.042 (-0.057;-0.028)	0.130 (0.059;0.201)
9	-0.124 (-0.142;-0.106)	0.114 (0.017;0.211)
10	-0.227 (-0.249;-0.205)	0.075 (-0.048;0.199)
11	-0.344 (-0.370;-0.318)	0.024 (-0.126;0.174)
12	-0.464 (-0.498;-0.431)	-0.031 (-0.208;0.146)
13	-0.582 (-0.629;-0.534)	-0.083 (-0.289;0.123)
14	-0.694 (-0.764;-0.625)	-0.129 (-0.367;0.108)
15	-0.803 (-0.903;-0.704)	-0.172 (-0.445;0.100)
16	-0.911 (-1.049;-0.773)	-0.214 (-0.527;0.098)
17	-1.020 (-1.204;-0.835)	-0.257 (-0.615;0.101)
18	-1.129 (-1.367;-0.890)	-0.299 (-0.710;0.111)

CIs at the 5% level are reported in parentheses.