Causal analysis of education and birth inequalities through a latent class structural equation model

Silvia Bacci, Francesco Bartolucci and Luca Pieroni

Abstract We investigate, in a causal perspective, whether the maternal education and other social characteristics, such as marital status, can influence birth inequalities, mainly birthweight and gestational age. We base the analysis on a structural equation model that takes explicitly into account the unobserved heterogeneity through a latent background variable, which accounts for certain types of confounder. The proposed model is a special case of finite mixture or latent class structural equation model, based on a suitable number of recursive equations, in which: (i) unobserved heterogeneity is represented by a discrete latent variable, (ii) maternal education and marital status may depend on the discrete latent variable and on other covariates, (iii) birthweight and gestational age depend on maternal education and marital status, on the discrete latent variable, and on other covariates. The study is based on the population of singleton born in Umbria (Italy) from 2007 to 2009.

Key words: EM algorithm, finite mixture models, ordinal response variables

1 Introduction

A large part of the economic literature ([1], [2], [3], [4]) found a strong correlation between maternal social characteristics, mainly education, and infant's health, gen-

Silvia Bacci
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli 20, 06123 Perugia e-mail: silvia.bacci@stat.unipg.it

Francesco Bartolucci
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli 20, 06123 Perugia e-mail: bart@stat.unipg.it

Luca Pieroni
Dipartimento di Economia, Finanza e Statistica, Università di Perugia, Via A. Pascoli 20, 06123 Perugia e-mail: lpieroni@unipg.it

erally measured as birthweight or gestational age. As health inequality at birth is a powerful determinant of health outcomes as adult, it is important to understand how much of the mentioned correlation is due to observed or unobserved confounding effects and how much is due to a causal relationship.

Several approaches to causal inference have been proposed in the literature (e.g., [5], [8], [10]). Here, we refer to the Pearl's approach [9], who extends the role of structural equation models [7] in the econometric literature to take into account the causal interpretation of the coefficients. Indeed, the partial regression coefficients can be appropriately interpreted in terms of causal effects on the response variable, given that *all* the relevant background variables have been included in the model. In this regard, we base our causal analysis on a model that takes explicitly into account the possible existence of unexplained heterogeneity through a latent background variable, which accounts for unobserved confounders. More precisely, with respect of standard structural equation models that are based on continuous observed and latent variables, our model is distinguished for two main elements:

1. a generalized linear modeling formulation [11] is adopted, so as to accommodate for mixed response types, that is both continuous and ordinal or binary observed responses, in the same set of structural equations;
2. the discreteness of the latent variable is assumed, in order to allow for (*i*) a semi-parametric estimation and (*ii*) detection of unobserved homogeneous classes of individuals who have similar latent characteristics.

The adopted model is a special case of mixture or latent class structural equation model [12], based on a suitable number of recursive equations, in which: (*i*) unobserved heterogeneity is represented by a discrete latent variable, (*ii*) the treatment variables may depend on the discrete latent variable and on other covariates, (*iii*) the response variables of interest depend on the treatment variables, on the discrete latent variable, and on other covariates.

The study is based on data obtained from the Standard Certificates of Live Birth (SCLB) collected in Umbria (Italy) in 2007, 2008, and 2009. For the study we limited our attention to natural conceptions, primiparous women, and singleton births; moreover, only infants with a gestational age of at least 23 weeks and a birthweight of at least 500 grams are taken into account. The total size of the sample, which merges each mother and her baby, amounts to 9005.

The outline of the article is as follows. Section 2 describes the proposed approach. Firstly, the theoretical background is illustrated and, then, its formulation in terms of a latent class regression model is shown. The main results are presented in Section 3.

2 The proposed latent class structural equation model

On the basis of the mentioned literature about causes of inequalities at birth and taking into account the information from SCLB, we assume that age and citizenship are

attributes of women that are not modifiable, education level may have a causal effect on marital status, and both education level and marital status may have a causal effect on gestational age and birthweight; moreover, we assume that gestational age and birthweight are inequality indicators with a likely high level of association, but without a specific causal relationship.

Coherently to this conceptual framework, our model is based on 4 equations, one for each endogenous variable. In order to take into account the possible presence of an unobserved confounding effect, we introduce, in the h-th equation referred to mother i, with $h = 1,\ldots,4$ and $i = 1,\ldots,n$, the discrete latent variable u_i. The distribution of each u_i is based on k support points, denoted by $\alpha_1^{(h)},\ldots,\alpha_k^{(h)}$, with mass probabilities $\pi_1,\ldots,\pi_k$. We assume that the k latent classes differ one another for different intercepts or cut-points (according to the type of response variable), wheres the functional form of each regression equation and the values of the structural coefficients are assumed to be constant among the classes.

We denote by $\mathbf{x}_i$ the vector of exogenous variables and by $\mathbf{z}_i$ and $\mathbf{y}_i$ the vectors of endogenous variables referred to mother i, with $i = 1,\ldots,n$, distinguishing the putative causes ($\mathbf{z}_i$) from the corresponding effects ($\mathbf{y}_i$). The vector $\mathbf{x}_i$ is composed by observations on variables *age* and *age*^{2}, both centered with respect to the sample mean, and on dummies referred to *citizenship* (Italian is the reference category). Vector $\mathbf{z}_i$ contains observations for individual i on *education level* (z_{1i}; reference category is middle school or less) and *marital status* (z_{2i}; reference category is married), whereas the vector $\mathbf{y}_i$ contains observations on *gestational age* (y_{1i}) and *birthweight* (y_{2i}). Moreover, vectors $\boldsymbol{\beta}^{(h)}$ and $\boldsymbol{\gamma}^{(h)}$ refer to the structural coefficients of covariates $\mathbf{x}_i$ and $\mathbf{z}_i$, respectively, and to identify the model, we let $\alpha_1^{(h)} = 0$, with $h = 1,\ldots,4$.

The first equation explains the *education level* as a function of the exogenous variables, through a global (or proportional odds) logit parametrization:

$$\log \frac{p(z_{1i} \geq j|\alpha_{u_i}^{(1)},\mathbf{x}_i)}{p(z_{1i} < j|\alpha_{u_i}^{(1)},\mathbf{x}_i)} = \nu_j^{(1)} + \alpha_{u_i}^{(1)} + \mathbf{x}_i'\boldsymbol{\beta}^{(1)} \quad j = 1,2, \tag{1}$$

where $j = 1$ for high school and $j = 2$ for degree or above. The second equation explains the *marital status* as a function of the exogenous variables and the *education level*, through a binary logit model, being $z_{2i} = 1$ for not married women and $z_{2i} = 0$ otherwise:

$$\log \frac{p(z_{2i} = 1|\alpha_{u_i}^{(2)},\mathbf{x}_i,\mathbf{z}_{1i})}{p(z_{2i} = 0|\alpha_{u_i}^{(2)},\mathbf{x}_i,\mathbf{z}_{1i})} = \nu^{(2)} + \alpha_{u_i}^{(2)} + \mathbf{x}_i'\boldsymbol{\beta}^{(2)} + \mathbf{z}_{1i}'\boldsymbol{\gamma}^{(2)}, \tag{2}$$

where $\mathbf{z}_{1i}$ indicates a vector of dummies for the variable z_{1i}. Finally, a bivariate normal regression model is specified for *gestational age* and *birthweight*, allowing for a conditional correlation different from zero:

$$\begin{cases} E(y_{1i}|\alpha_{u_i}^{(3)},\mathbf{x}_i,\mathbf{z}_i) = \nu^{(3)} + \alpha_{u_i}^{(3)} + \mathbf{x}_i'\boldsymbol{\beta}^{(3)} + \mathbf{z}_i'\boldsymbol{\gamma}^{(3)}, \\ E(y_{2i}|\alpha_{u_i}^{(4)},\mathbf{x}_i,\mathbf{z}_i) = \nu^{(4)} + \alpha_{u_i}^{(4)} + \mathbf{x}_i'\boldsymbol{\beta}^{(4)} + \mathbf{z}_i'\boldsymbol{\gamma}^{(4)}. \end{cases} \tag{3}$$

For the selection of the optimal number k of latent classes we suggest to use the Bayesian Information Criterion (BIC). The proposed model is estimated through an Expectation Maximization algorithm [6] we implemented in a set of R functions.

3 Main results

For the data at hand, on the basis of BIC we selected $k = 3$ latent classes with weights equal to 0.931, 0.041 and 0.028. Women in class 1 represent the main part of population and they are characterized, in average, by a gestational age equal to 39.5 weeks and infants with a weight equal to 3.2 Kg. Women in class 2 deliver 6.3 weeks before and their infants are 1.3 Kg. lighter. Finally, class 3 is characterized by women whose infants are significantly heavier (0.7 Kg.), although no significant difference in gestational age results.

As concerns the causal pathways, we firstly observe that, while mother education has a positive causal effect on a mother's probability of being married, the marital status does not appear to be a significant determinant to explain the differences in birth outcomes. Secondly and most important, after having controlled for unobserved confounders, birthweight is positively influenced by mother's education, whereas gestational age is not affected.

References

1. Abrevaya, J., Dahl, C. M.: The effects of birth inputs on birthweight. J. Bus. Econ. Statist. **26**, 379 – 397 (2008)
2. Almond, D., Chay, K. Y., Lee, D. S.: The costs of low birth weight. Q. J. Econ. **120(3)**, 1031 – 1083 (2005)
3. Currie, J.: Inequality at birth: some causes and consequences. Am. Econ. Rev. **101(3)**, 1 – 22 (2011)
4. Currie, J., Moretti, E.: Mother's education and the intergenerational transmission of human capital: evidence from college openings. Q. J. Econ. **118(4)**, 1495 – 1532 (2003)
5. Dawid, A.: Influence diagrams for causal modelling and inference. Int. Stat. Rev. **70**, 161 – 189 (2002)
6. Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Stat. Soc. B Met., **39**, 1–38 (1977)
7. Duncan, O.: Introduction to structural equation models. Academic Press, New York (1975)
8. Lauritzen, S.: Graphical models. Claredon Press, Oxford (1996)
9. Pearl, J.: Causality: models, reasoning, and inference. Cambridge University Press, New York (2000)
10. Rubin, D.: Estimating causal effects of treatments in randomized and non randomized studies. J. Educ. Psychol. **66**, 688 – 701 (1974)
11. Skrondal, A., Rabe-Hesketh, S.: Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC (2004)
12. Vermunt, J., Magidson, J.: Structural Equation Models: mixture models. In: Encycopledia of Statistics in Behavioral Science, pp. 1922–1927. Wiley (2005)