A causal analysis of mother's education on birth inequalities

Silvia Bacci*[1], Francesco Bartolucci*, Luca Pieroni*

*Dipartimento di Economia, Finanza e Statistica - Università di Perugia

Università La Sapienza, Roma, 20-22 June 2012

[1] silvia.bacci@stat.unipg.it

Outline

1. Introduction
    - Background
    - Aim and method
    - The dataset
2. The theoretical model
3. Preliminary analyses
4. Methodological aspects
    - Causal analysis
    - Structural Equations Models and extensions
5. The proposed finite mixture SEM
6. Main results
7. Conclusions

Introduction

- Motivation:

  The actual benefits of any public health initiative aimed at reducing health inequality at birth crucially depend upon the estimates of the causal effect of mother's characteristics and the possibility of intervention by policy-makers

- Aim:

  Investigating about the causal relation between mother's social characteristics and infant's health

Background

Several classical economical references analyze the impact of maternal social characteristics and behaviors on infant health:

- a strong correlation was found between mother’s education and birthweight
- lacks on mother’s education may yield effects on the initial endowment of an infant’s health and it tends to be pervasive over the life
- the initial inequality may partly be transmitted from a generation to the next, with the effect of a lower educational attainment, poorer health status, and reduced earning in adult age
- References: Rosenzweig and Schultz (1983), Rosenzweig and Wolpin (1991), Currie and Moretti (2003)

Aim

- Our aim is to investigate about the causal effect of maternal social characteristics, such as education and marital status, on birth inequality outcomes measured by gestational age and birthweight
- We refer to the Pearl's approach to causal inference, based on Structural Equation Models (SEMs)
- We account for unobserved heterogeneity (or confounding) by introducing a discrete latent background variable

Method

The proposed methodological approach is a special case of finite mixture SEM based on a suitable number of consecutive equations in which:

1. unobserved heterogeneity is represented by a discrete latent variable defining latent classes of individuals
2. the causes may depend on the discrete latent variable and on other covariates
3. the response variables of interest depend on the causes, on the discrete latent variables, and on other covariates

In this way, since the causal effect is evaluated within homogenous groups of individuals, it is still possible to read the partial regression coefficients in terms of causal effects, as it happens when we adjust for observed confounders

The dataset

- data are collected in Umbria (Italy) in years 2007, 2008, 2009
- data come from the Standard Certificates of Live Birth (SCLB)
- SCLB contain socio-economic and demographic information on mothers and their infants
- our study is focused on a subset of 9005 records corresponding to (i) natural conceptions, (ii) primiparous women, (iii) singleton births, (iv) infants with a gestational age of at least 23 weeks and a birthweight of at least 500 grams

Descriptive analysis

Table: Distribution of variables

Variable	Category	%	Mean	St.Dev.
Gestational age (weeks)			39.310	1.686
Birthweight (kg)			3.262	0.487
Age (years)			30.040	5.288
Citizenship	Italian	80.1		
	east-Europe	12.6		
	other citizenship	7.3		
Education level	middle school or less	19.8		
	high school	51.9		
	degree and above	28.4		
Marital status	married	70.0		
	not married	30.0		

The theoretical model

We assume that

- gestational age and birthweight are inequality indicators with a likely high level of correlation but without a specific causal relationship
- age and citizenship are attributes of mothers that are not modifiable
- educational level may have a causal effect on marital status
- both marital status and educational level may have a causal effect on gestational age and birthweight

Notation

- $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ is the vector of birth outcomes (gestational age, birthweight) for each singleton deliver i, $i = 1, \ldots, n$
- $\boldsymbol{z}_i = (z_{i1}, z_{i2})$ is the vector of putative causes (mother education, marital status)
- $\boldsymbol{x}_i$ is a vector of mother-specific not modifiable characteristics (citizenship, age) other than those included in $\boldsymbol{z}_i$
- $\boldsymbol{u}_i$ reflects mother-specific unobservable determinants of child outcomes (e.g., genetic factors, unreported life style behaviors)

Multiple regressions

Table: Regression for the gestational age

covariate	category	est.	s.e.	t stat.	p-value
intercept	–	39.325	0.051	772.686	0.000
age	–	-0.019	0.004	-4.910	0.000
age^2	–	-0.001	0.001	-1.336	0.181
citizenship	Italian	0.000	–	–	–
citizenship	east-Europa	-0.242	0.059	-4.099	0.000
citizenship	other citizenship	-0.208	0.072	-2.887	0.004
education	middle school or less	0.000	–	–	–
education	high school	0.077	0.049	1.551	0.121
education	degree or above	0.077	0.057	1.345	0.179
marital	married	0.000	–	–	–
marital	not married	-0.025	0.039	-0.640	0.522

Multiple regressions

Table: Regression for the birthweight

covariate	category	est.	s.e.	t stat.	p-value
intercept	–	3.240	0.015	220.413	0.000
age	–	-0.005	0.001	-4.159	0.000
age^2	–	-0.000	0.000	-0.875	0.381
citizenship	Italian	0.000	–	–	–
citizenship	east-Europa	0.032	0.017	1.847	0.065
citizenship	other citizenship	-0.050	0.021	-2.414	0.016
education	middle school or less	0.000	–	–	–
education	high school	0.032	0.014	2.243	0.025
education	degree or above	0.050	0.017	3.033	0.002
marital	married	0.000	–	–	–
marital	not married	-0.019	0.011	-1.682	0.092

Confounding effect

Confounding effect: when two variables z and y have a common cause u that confounds the true relationship between the putative cause z and the effect y (case (a))

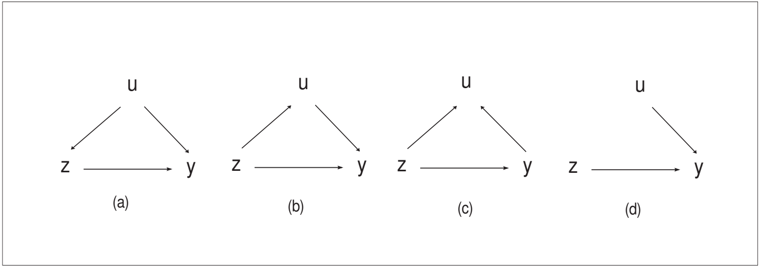

Figure: *Causal relation between z and y and presence of a third variable u: (a) u as common cause, (b) u as intermediate effect, (c) u as common effect, (d) u as cause acting independently from z*

SEM-based approach

- an useful instrument to control for confounding bias is represented by SEMs
- the partial regression coefficients of a SEM 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
- unfortunately, after having controlled for the observed covariates, the residual unexplained heterogeneity may be still substantial . . .

Extensions of standard SEM

Finite Mixture SEM:

- we assume that the unobserved heterogeneity may be captured by a limited number K of (unobserved) groups or classes of individuals
- the K latent classes differ one another for different intercepts, while the functional form of each regression equation and the values of structural coefficients are assumed to be constant among the classes
- Advantages of finite mixture SEM:
  - each mixture component identifies homogeneous classes of individuals that have very similar latent characteristics, so that, in a decisional context, individuals in the same latent class will receive the same treatment
  - the model estimation does not require any parametric assumption on the latent variable distribution

Extensions of standard SEM

Mixed types of response:

To accomodate continuous, ordinal, and binary responses we introduce a latent continuous variable z^*_{il} underlying each observable response variable z_{il}

$$z_{il} = G_l(z^*_{il})$$

where $G_l(\cdot)$ is defined according to the different nature of z_{il}:

1. when the observed response is of a continuous type, an identity function is adopted $G_l(z^*_{il}) = z^*_{il}$
2. when the observed response is binary, then $G_l(z^*_{il}) = I\{z^*_{il} > 0\}$
3. when the observed response is ordinal with categories $j = 1, \dots, J_l$, we introduce a set of cut-points $\tau_{l1} \geq \dots \geq \tau_{l,J_l-1}$ and we define

$$G_l(z^*_{il}) = \begin{cases} 1 & z^*_{il} \leq -\tau_{l1}, \\ 2 & -\tau_{l1} < z^*_{il} \leq -\tau_{l2}, \\ \vdots & \vdots \\ J & z^*_{il} > -\tau_{l,J_l-1} \end{cases}$$

The proposed model

- Equation 1 (educational level):

  $z_{i1} = G_1(z^*_{i1})$, with G_1 defined as in (3) and

  $$z^*_{i1} = \mu_1 + \alpha_{i1} + \boldsymbol{x}'_i\boldsymbol{\beta}_1 + \varepsilon_{i1}$$

  - $\mu_1 + \alpha_{i1}$ is a specific intercept for subject i
  - $\boldsymbol{\beta}_1$ is a vector of regression coefficients for the covariates in $\boldsymbol{x}_i$
  - ε_{i1} is a random error term with logistic distribution

- Equation 2 (marital status):

  $z_{i2} = G_2(z^*_{i2})$, with G_2 defined as in (2) and

  $$z^*_{i2} = \mu_2 + \alpha_{i2} + \boldsymbol{x}'_i\boldsymbol{\beta}_2 + z'_{i1}\gamma + \varepsilon_{i2}$$

  - $\mu_2 + \alpha_{i2}$ is the subject specific intercept
  - $\boldsymbol{\beta}_2$ and γ are regression coefficients
  - ε_{i2} is an error term with logistic distribution, which is independent of ε_{i1}

- Equation 3 (gestational age, birthweight):

$$\boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\delta}_i + \boldsymbol{\Phi}\boldsymbol{x}_i + \boldsymbol{\Psi}\boldsymbol{z}_i + \boldsymbol{\eta}_i$$

  - $\boldsymbol{\nu} = (\nu_1, \nu_2)'$; $\boldsymbol{\delta}_i = (\delta_{i1}, \delta_{i2})'$; $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)'$; $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2)'$;
  - $\boldsymbol{\eta} = (\eta_1, \eta_2)'$ is a vector of error terms, following a bivariate Normal distribution centered at $\mathbf{0}$ and with variance-covariance matrix $\boldsymbol{\Sigma}$ and independent of the ε_{i1} and ε_{i2}
  - $\nu_1 + \delta_{i1}$ is the subject-specific intercept for the gestational age
  - $\nu_2 + \delta_{i2}$ is the subject-specific intercept for the birthweight
  - $\boldsymbol{\phi}_1$ and $\boldsymbol{\psi}_1$ are the regression coefficients for the first response variable
  - $\boldsymbol{\phi}_2$ and $\boldsymbol{\psi}_2$ are the regression coefficients for the second response variable

Note that $\alpha_{i1}, \alpha_{i2}, \boldsymbol{\delta}_i$ have a discrete distribution with K support points and corresponding weights

The proposed model may be estimated by the maximum likelihood method, efficiently implemented through an EM algorithm

Model selection

A crucial point with mixture models concerns the choice of the number k of mixture components

- coherently with the main literature we suggest to use BIC index

$$\mathrm{BIC} = -2\hat{\ell} + \log(n)\#\mathrm{par}$$

- we fit the finite mixture SEM with increasing K values, relying the choice of optimal K on the value just before the first increasing of the BIC index
- we obtain the minimum BIC value in correspondence of $K = 3$ latent classes

K	$\hat{\ell}$	#par	BIC
1	-35700.768	32	71692.914
2	-34536.422	37	69409.750
3	-34488.589	42	69359.610
4	-34467.548	47	69363.055

Regression results about education

covariate	category	est.	s.e.	t stat.	p-value
intercept (μ_1)	–	2.053	0.039	52.285	0.000
1st cutpoint (τ_1)	–	0.000	–	–	–
2st cutpoint (τ_2)	–	-2.695	0.031	-20.780	0.000
age	–	0.103	0.004	23.405	0.000
age^2	–	-0.009	0.001	-14.587	0.000
citizenship	Italian	0.000	–	–	–
citizenship	east-Europa	-0.806	0.069	-11.712	0.000
citizenship	other citizenship	-1.100	0.086	-12.780	0.000

Regression results about marital status

covariate	category	est.	s.e.	t stat.	p-value
intercept (μ_2)	–	-0.763	0.065	-11.313	0.000
age	–	-0.027	0.005	-5.487	0.000
age^2	–	0.008	0.001	12.381	0.000
citizenship	Italian	0.000	–	–	–
citizenship	east-Europa	-0.679	0.082	-8.264	0.000
citizenship	other citizenship	-0.677	0.101	-6.701	0.000
education	middle school or less	0.000	–	–	–
education	high school	-0.152	0.064	-2.375	0.018
education	degree or above	-0.468	0.076	-6.123	0.000

Regression results for gestational age

covariate	category	est.	s.e.	t stat.	p-value
intercept (ν_1)	–	39.346	0.044	905.935	0.000
age	–	-0.015	0.003	-4.789	0.000
age^2	–	-0.001	0.000	-2.544	0.011
citizenship	Italian	0.000	–	–	–
citizenship	east-Europa	-0.194	0.049	-3.942	0.000
citizenship	other citizenship	-0.112	0.060	-1.855	0.064
education	middle school or less	0.000	–	–	–
education	high school	0.025	0.042	0.608	0.543
education	degree or above	0.029	0.049	0.600	0.548
marital	married	0.000	–	–	–
marital	not married	0.025	0.033	0.749	0.454

Regression results for birthweight

covariate	category	est.	s.e.	t stat.	p-value
intercept (ν_2)	–	3.238	0.017	195.392	0.000
age	–	-0.004	0.001	-3.863	0.000
age^2	–	-0.000	0.000	-1.708	0.088
citizenship	Italian	0.000	–	–	–
citizenship	east-Europa	0.041	0.016	2.653	0.008
citizenship	other citizenship	-0.031	0.019	-1.608	0.108
education	middle school or less	0.000	–	–	–
education	high school	0.023	0.014	1.674	0.094
education	degree or above	0.043	0.017	2.462	0.014
marital	married	0.000	–	–	–
marital	not married	0.011	0.012	0.904	0.366

Latent structure

Table: Support points and class weights estimates

	$k = 1$	$k = 2$	$k = 3$
education	0.005	-0.165 (0.234)	-0.005 (0.964)
marital status	0.026	0.289 (0.081)	-0.794 (0.021)
gestational age	0.178	-6.086 (0.000)	0.123 (0.671)
birthweight	0.005	-1.245 (0.000)	0.728 (0.000)
class weight (π_k)	0.931	0.028	0.041

- women from class 1 represent the main part of the population (93.1%)
- women from class 2 present a significant higher propensity (at 10% level) to be not married and to give birth 6.1 weeks before; their infants weigh 1.245 kg less; no significant difference results about the educational level
- women in class 3 have a higher tendency to be married and the birthweight of their infants is significantly higher ($+0.728$ kg); no significant difference results with respect to educational level and gestational age

Main conclusions about causal effects

- about the marital status, the analysis confirms the absence of any causal effect on both gestational age and birthweight
- about the educational level, results suggest a significant and positive effect of education on the probability to be married
- about the educational level, the increase of p-values denote the presence of a confounding effect on both gestational age and birthweight
- however, even after controlling for a latent common cause, a significative effect persists on the birthweight: a higher educational level causes a higher birthweight
- our interpretation of this result is that the woman's educational level is related with specific unobservable variables, such as the ability to properly manage the pregnancy so as to improve the health level of the newborn
- our result confirm that improving education among young mothers should be viewed as a key policy to reduce costs of unhealthy child outcome