Società italiana di economia pubblica

XIX CONFERENZA

ECONOMIA DEL CAPITALE UMANO
Istituzioni, incentivi e valutazioni

Pavia, Aule storiche Università, 13 - 14 settembre 2007

ROLE EFFECTS IN EDUCATION TRANSMISSION

VALENTINO DARDANONI, ANTONIO FORCINA, AND SALVATORE MODICA

pubblicazione internet realizzata con contributo della COMPAGNIA di San Paolo

ROLE EFFECTS IN EDUCATION TRANSMISSION

VALENTINO DARDANONI, ANTONIO FORCINA, AND SALVATORE MODICA

ABSTRACT. Recent literature has analyzed the impact of parents' education on children's by controlling for parents' unobservable endowments. By applying recent advances in latent class analysis as a tool for modeling individual heterogeneity, we study the direct causal effect of parents' education conditional on *children's* unobservable schooling endowment. We interpret this effect as reflecting social-norms based parental pressure, and we call it a *role effect.* In the UK NCDS dataset that we use we find that this effect is sizeable, and by looking separately at sons and daughters subsamples it emerges that it is mostly confined to the father–son relationship.

JEL Classification Numbers I21, C35

Keywords Education Transmission, Unobserved Endowments, Finite Mixtures, Direct Causal Effect.

1. INTRODUCTION

The present research is concerned with the widely studied problem of measuring the effect of family background on educational achievement, taking into account individual unobserved heterogeneity. In particular we are interested in the role of mothers and fathers in education transmission, upon which contrasting evidence reported in recent studies (see e.g. the survey by Holmlund–Lindahl–Plug [26], more below) calls for reflection.

It is by now well appreciated that finding a positive association between parents' and children's education level may simply reflect correlation between underlying unobserved heritable endowments: better endowed parents are more educated and have better endowed, hence more educated, children. To discuss the causal relations between parents' and children's unobservable endowments U^p, U^c and their schooling S^p, S^c it may be useful to start with a simple system of linear equations:

$$\begin{aligned} S^c &= a + bS^p + cU^c + \epsilon \qquad (1) \\ U^c &= d + eS^p + fR^p + gU^p + \eta\,, \end{aligned}$$

Date: January 30, 2007.

where R^p denotes the parent's child-rearing ability; we consider here a single parent for illustration, and ϵ, η are assumed uncorrelated with the covariates. The first equation says that a person's education depends on her schooling endowment and her parents' education. In fact, *given her own endowment* U^c, it is not obvious why parents' education should enter the function as an independent argument, and we shall shortly discuss this point. The second equation can be considered a typical nature–nurture relation; upon substitution in the above equations one obtains the standard reduced form equation (compare equation (2) in [7] or equation (9) in [26]):

$$S^c = \alpha + \beta S^p + \gamma R^p + \delta U^p + \psi. \tag{2}$$

Equation (2) shows that the marginal causal effect of S^p on S^c, namely $\beta = b + ce$, can be decomposed into a component which captures the impact of parents' schooling on children's endowments, and, if $b \neq 0$, a direct effect independent of endowments transmission. Clearly estimation of the marginal causal effect (β in equation (2)) requires controlling for U^p; on the other hand, estimation of the causal effect of S^p conditional on the child's endowment (b in equation (1)) requires controlling for U^c. This is apparent in the above linear equations, but as we show in the Appendix (section 6.1) it continues to hold in the nonlinear context of discrete response variables which will be the setting of the present paper.

Estimation of the effect of parents' education after controlling for their unobservable endowments has been the object of recent research initiated by Behrman and Rosenzweig [7], who challenge conventional wisdom (cfr. the classical surveys [6, 24]) that parents' schooling has significant effect on their children's (and that generally mothers' schooling has a greater effect than fathers'), on the grounds that, without proper control for unobservable endowments, resulting estimates may be biased and reflect correlation but not causation. Behrman and Rosenzweig arrive at the striking result that mothers' education has no effect on children's after controlling for parents' endowments by taking differences on MZ twin parents.[1] In two important follow-ups of Behrman–Rosenzweig [7], Plug [31] using adoptees confirms the finding that

[1] To illustrate, under the assumption that MZ twins' endowments *and* child-rearing abilities are equal, denoting twins by subscripts 1 and 2 we have $U_1^p = U_2^p$ and $R_1^p = R_2^p$, whence by taking differences in equation (2) above one gets $S_1^c - S_2^c = \beta(S_1^p - S_2^p) + \varphi$, from which an unbiased estimate of β can be obtained if $(S_1^p - S_2^p) \neq 0$ for a sufficient number of subjects.

only the father's education has a positive impact on the child's, while Black–Devereux–Salvanes [11], using reforms in municipal compulsory schooling laws as instruments, find almost no causal link between parents' and children's education.[2]

The difficulties in controlling for parents' unobservable endowments are illustrated in the survey by Holmlund–Lindahl–Plug in [26], where these three different methods (namely use of twins, adoptees and schooling laws instruments) are applied to a single data set, and it is shown that the three approaches produce results which are in conflict with each other.[3] The fact that estimates of β may be quite sensitive to the key assumptions made in the process is compounded with the problem that separate estimation of fathers' and mothers' effects requires control for assortative mating.

In this paper we consider the causal effect of parents' education on their children's after controlling for *children*'s unobservable schooling endowment (the analogy with the simple linear model (1) and (2) would be that we estimate b rather than β). In the terminology of the literature on causality (see e.g. Pearl [30] page 126) this is called a direct effect. For this purpose we consider as dependent variable, rather than years of schooling, an indicator of *schooling attainment* in terms of achievement of a significant educational certification, since this is more likely to reflect the value assigned to education by the subjects and their parents. In the context of the British educational system, such a certification is represented by the O-Level exams, which are passed by about 50% of our sample of students born in UK in 1958 (see section 2 below). By exploiting the unique features of the English NCDS dataset, which contains information on a rich set of education related variables taken from very early age, we identify the unobserved children's schooling endowment U^c by using a finite mixture model (the implied meaning of U^c is described on page 10 below).

The natural interpretation of the effect we look at emerges by asking why, among subjects with a given level of schooling endowment, those with more

[2] On Behrman and Rosenzweig see also the critical Comment [1] by Antonovics and Goldberger (and the authors' Reply [8]). Intergenerational education transmission is also analyzed by Bjorklund–Lindahl–Plug [10], Dearden–Machin–Reed [17] and Sacerdote [32], [33] using adoptees, and by Chevalier [15] and Oreopoulus–Page–Stevens [29] using schooling laws instruments.

[3] "The twin and adoption methods give us positive intergenerational schooling coefficients for fathers but no effects for mothers. Whereas the IV strategy indicates that it is only the mother's education that is important", [26], p.51.

educated parents should attain higher educational levels. The answer may be that they have stronger *social motivation*, which is what the effect reflects. The effect may include, for example, the intergenerational transmission of education-dependent labor markets skills, better information on the value of education, or simply greater parental pressure reflecting social norms. We call it a *role effect*. In any given social context, note that this kind of influence is typically gender dependent, in the sense that mothers and fathers may have different effects on daughters and sons.[4]

Our econometric analysis exploits recent advances in the theory of marginal modelling (see e.g. Bergsma and Rudas [5]), which allow identification and estimation of finite mixture models where not only response variables may depend on covariates, but also some residual association between them is permitted. The actual regression system we estimate is displayed in (5) on page 12 below.

The empirical findings we report confirm the presence of role effects in our sample: given the child's schooling ability, more educated parents bring their offsprings to a greater level of education. In fact, in the pooled sample of sons and daughters only fathers' education is found to have a positive impact on children's. But the nature of the effect we estimate appears more neatly by examining its gender dependence: when the sons and daughters samples are analyzed separately, it is seen that fathers' education affects *only* that of their sons; mothers' education has a weak (and not significant) effect, only on daughters. It may be worth noting that this strong father–son link, which reflects the social-norms based pressure that families put on children for educational attainment, may well be confined to the social context to which our data refers. We discuss these findings in the conclusions after we present estimation results.

The rest of the paper is organized as follows. The data are described in section 2; we then derive the model to be estimated (section 3) and report the results of estimation (section 4). Section 5 contains some concluding remarks. In the Appendix we explain the details of the identification and estimation of our model using likelihood inference techniques.

[4] In studies of educational attainment, early mention of the possibility of interclass differences in educational choice at given levels of academic performance is found in Boudon [13], who called this a 'secondary effect'. Erikson et al. [20] confirm the presence of this effect by counterfactual analysis.

2. The Data Structure

In the British educational system, students at the age of 16 take the so called O-Level exams on a set of chosen topics. If a student has reached a minimum standard in terms of quantity of subjects taken and grades obtained, she is awarded an O-Level certification and allowed, if she wishes, to access the next level of education (the so-called A-Level).

We use data from the National Child Development Survey (NCDS). This data set is a UK cohort study targeting all the population born in the UK between the 3rd to the 9th of March 1958. Individuals were surveyed at different stages of their life and information on their schooling results and their background was collected. Our main dependent variable is the binary variable OL which takes value 1 iff the subject has obtained the necessary formal O-Level qualifications; in our sample about fifty percent of the subjects achieve them.[5] [6]

A rather unique feature of the NCDS is that subjects were tested for reading and mathematical ability at the early age of seven and then again at eleven and sixteen. These math and reading test score results can be used for identification of the unobservable endowment U^c. Here we use binary response variables extracted from the cognitive tests which act as multiple indicators for U^c, with the dichotomized variables taking value 1 iff the subject's score is at or above the sample median. These binary variables will be denoted as $M_{16}, R_{16}, M_{11}, R_{11}, M_7, R_7$, where M, R stand for math and reading and subscripts denote the age at which the tests are taken.

Parents' schooling is defined as the age at which they left school; fathers' and mothers' schooling are collected into the vector $\boldsymbol{x}$. Regarding other family background variables, the NCDS contains also information on parents' interest in their child's education, as reported by teachers separately for mothers and fathers; this can be considered a proxy for child-rearing ability R^p. Data on parents' interest are originally classified into 5 distinct categories (overconcerned, very interested, shows some interest, little interest, can't say); we have chosen to group them into two (interested in the first three cases and not interested in the last two), so that a parent's interest is indicated by a

[5] Since attainment of O-Level certification is the first significant schooling continuation decision made by UK students, considering this as the dependent variable avoids the dynamic selection bias problem pinned down by Cameron-Heckman [14].

[6] The same data set is extensively used, among others, by Blundell–Dearden–Sianesi [12], who study the effect of education on earnings.

single dummy. To allow for possible capital market imperfections, we include also family income in the set of family background variables. The two parents' interest variables and family income are collected in the vector $\boldsymbol{z}$; the vector of all five family background characteristics is denoted by $\boldsymbol{b} = (\boldsymbol{x}, \boldsymbol{z})$.[7]

From the NCDS we selected all subjects for whom we had information on OL, test scores and $\boldsymbol{b}$; the resulting sample is made of 4553 subjects, 2308 males and 2245 females. Summary statistics on the data used are reported in Table 1 below. A complete description of the data is available at `http://www.esds.ac.uk/longitudinal/access/ncds`.

	Avg daughters	Avg sons	short name	type
O-Level	0.5465	0.4861	OL	dummy
Math at 16	0.4980	0.5750	M_{16}	dummy
Reading at 16	0.5416	0.5823	R_{16}	dummy
Math at 11	0.5376	0.5290	M_{11}	dummy
Reading at 11	0.5225	0.5199	R_{11}	dummy
Math at 7	0.4450	0.4900	M_7	dummy
Reading at 7	0.5889	0.4870	R_7	dummy
Father schooling	14.9327	14.8930	fs	numerical
Mother schooling	14.9871	14.9302	ms	numerical
Father interest	0.6735	0.6720	fi	dummy
Mother interest	0.7430	0.7132	mi	dummy
Family income	16.7995	16.7093	ty	numerical

Table 1. Data

3. Unobserved endowments and finite mixture models

3.1. **Preliminary remarks, notation and examples.** The finite-mixture approach is well known and much used in many branches of statistics such as biometrics and psychometrics (see e.g. [28, 35]). An early use in economics is in Heckman and Singer [25]; recently, it has been mainly used in dynamic models, with several applications to scholastic achievements. Arcidiacono and Jones [2] contains an ample discussion of this literature.

Consider a set of k binary response variables Y_j, $j = 1, \ldots, k$, taking values $y_j \in \{0, 1\}$, where, as usual, lower case letters denote observed values of the corresponding capital-letter random variables. A given response configuration

[7] We also considered parents' age, but dropped it from the analysis because it was never found significant.

will be denoted by the column vectors $\boldsymbol{y} = (y_1, \ldots, y_k)'$ and $q_{\boldsymbol{y}}$ will denote the probability of a given response configuration. Now order the response patterns lexicographically, with elements on the right running faster from 0 to 1. A probability distribution on the set of the 2^k distinct response configurations will be represented by the vector $\boldsymbol{q}$ having elements $q_{\boldsymbol{y}}$ ordered as above; this vector belongs to the simplex Δ_{2^k}.

The following simple example introduces some elementary ideas which will hopefully clarify the approach we take in this paper. Let $\boldsymbol{q} \in \Delta_4$ denote the distribution of two binary response variables Y_1 and Y_2. Since probabilities add to one, we can describe $\boldsymbol{q}$ by an appropriate choice of 3 free *parameters*, where a parameter is any real valued function of $\boldsymbol{q}$. For example, $\boldsymbol{q}$ could be defined by two parameters describing the univariate marginals and a parameter describing their association such as $\Pr(Y_i = 1)$, $i = 1, 2$ and $\Pr(Y_1 = 1, Y_2 = 1)$. However, this parameterization is not unique; any *invertible function* of these 3 parameters can be considered equivalent, and depending on the purpose at hand alternative parameterizations may be preferable. For example, using 2 *logits* to describe the univariate marginals

$$\lambda_{Y_1} = \log\left[\frac{\Pr(Y_1 = 1)}{\Pr(Y_1 = 0)}\right], \ \lambda_{Y_2} = \log\left[\frac{\Pr(Y_2 = 1)}{\Pr(Y_2 = 0)}\right],$$

and a second order logit interaction parameter (also called *log-odds ratios*) to describe the bivariate association

$$\begin{aligned} \lambda_{Y_1,Y_2} &= \log\left[\frac{\Pr(Y_1 = 0, Y_2 = 0)\Pr(Y_1 = 1, Y_2 = 1)}{\Pr(Y_1 = 1, Y_2 = 0)\Pr(Y_1 = 0, Y_2 = 1)}\right] \\ &= \lambda_{Y_1|Y_2=1} - \lambda_{Y_1|Y_2=0} = \lambda_{Y_2|Y_1=1} - \lambda_{Y_2|Y_1=0} \end{aligned}$$

gives an invertible mapping from Δ_4 to $\mathbb{R}^3$, and thus can be considered as an equivalent alternative parameterization of $\boldsymbol{q}$ since it conveys all relevant information on the joint distribution of Y_1, Y_2 (see e.g. Bartolucci–Colombi–Forcina [3] for a general recursive definition of logit parameters and a discussion on their invertibility properties).

The above marginal parameterization is not unique; for comparison, consider the following recursive model: $\lambda_{Y_1} = a_1$ and $\lambda_{Y_2|Y_1} = a_2 + a_3 Y_1$. The 3 parameters a_1, a_2, a_3 form again an invertible mapping from Δ_4 to $\mathbb{R}^3$; notice also that by the definitions above $a_2 = \lambda_{Y_1,Y_2}$, Assuming that association parameters are zero clearly implies that $\boldsymbol{q}$ belongs to a subset of Δ_4; for example, if we assume that Y_1 and Y_2 are independent:

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1)\Pr(Y_2 = y_2),$$

then any $\boldsymbol{q} \in \Delta_4$ which satisfies the above relation can be uniquely described by two parameters such as $\Pr(Y_1 = 1)$, $\Pr(Y_2 = 1)$ or equivalently λ_{Y_1} and λ_{Y_2}.

Finally, suppose that Y is an *ordinal variable* taking values $y \in \{0, 1, \ldots, t\}$. A natural generalization of the binary logits above is obtained (exploiting the ordered nature of the support of Y) by the so called *global logits* $\gamma_Y^1, \ldots, \gamma_Y^t$:

$$\gamma_Y^y = \log\left[\frac{\Pr(Y \geq y)}{\Pr(Y < y)}\right], \quad y = 1, \ldots, t.$$

It can be seen that global logits are equivalent to standard logits under successive dichotomization of the support of Y; notice also that the set of global logits $\gamma_Y^1, \ldots, \gamma_Y^t$ forms an invertible parameterization of the distribution of Y under the assumption that $\gamma_Y^1 \geq \cdots \geq \gamma_Y^t$.

3.2. **Classical latent class analysis.** Given observed binary responses $(Y_1, \ldots, Y_k)$, classical latent class analysis (e.g. Goodman [21]) tries to identify a discrete random variable U taking values in $\{0, 1, \ldots, m\}$ such that

$$\begin{aligned}&\Pr(Y_1 = y_1, \ldots, Y_k = y_k) = \\ &\qquad \sum\nolimits_u \Pr(U = u)\Pr(Y_1 = y_1 \mid U = u) \cdot \ldots\ldots \cdot \Pr(Y_k = y_k \mid U = u); \quad (3)\end{aligned}$$

that is, the unobservable latent variable U makes observed responses conditionally independent. Clearly this assumption (which is known in the latent class literature as *local independence*) restricts the dimension of the probability space of $(U, Y_1, \ldots, Y_k) \equiv (U, \boldsymbol{Y})$ from $(m+1)\cdot 2^k - 1$ to $m + (m+1)\cdot k$. Indeed, any $\boldsymbol{p} \in \Delta_{(m+1)\cdot 2^k}$ which satisfies relation (3) is uniquely determined by the following parameters: $\Pr(U \geq u)$ for $u = 1, \cdots, m$, and $\Pr(Y_i = 1 \mid U = u)$ for $i = 1, \ldots, k,\ u = 0, \ldots, m$. Notice however that, since U is not observed, the dimension of the space of the observed responses is only $2^k - 1$, and this restricts the number of classes of U which can be identified, since $m + (m+1)\cdot k$ must be less than or equal to $2^k - 1$.

3.3. **The model estimated in this paper.** Since the child's endowment will be the only latent variable which we model explicitly, we will omit the superscript c from U^c. To identify the unobservable child's endowment U one can extract information from response vectors which act as *multiple indicators* of the latent variable U and, as in finite mixture models, we will assume that U is a discrete random variable taking values in $\{0, 1, \ldots, m\}$.

As indirect indicators of schooling ability we shall use dichotomized observations on maths and reading test scores. To be more specific note that in the case at hand the unobservable residual heterogeneity affecting a given educational attainment (the English O-Level certification in our case) can be seen as the result of two different factors: early schooling endowments and relevant knowledge acquired through learning. Since this is the individual heterogeneity which our latent variable U should capture, we shall have it *jointly* identified, besides OL itself, by early indicators of innate mathematical and reading comprehension on the one hand, and by indicators of the level of acquired knowledge on the other.

As early indicators we take performance in math and reading tests taken at 7 and 11 years of age; as indicators of OL-relevant knowledge we have performance in math and reading tests taken at 16 (approximately the same year of OL exams). So in our case $k = 7$ and $\boldsymbol{Y} = (OL, M_{16}, R_{16}, M_{11}, R_{11}, M_7, R_7)$. If a random variable U which satisfies (3) could be identified, intuitively U would capture underlying unobserved cognitive ability, since knowledge of U would imply for example that knowledge of any test result would be irrelevant for predicting OL results and viceversa.

However, classical latent class models may be too restrictive in our specific context, in particular the assumptions underlying equation (3). The first point to notice is that it seems plausible that endowments and responses are themselves affected by family background characteristics. Thus, a first extension of the classical model is to allow the distribution of the response vector $(U, \boldsymbol{Y})$ to depend on observable covariates, which we recall are denoted by $\boldsymbol{b} = (\boldsymbol{x}, \boldsymbol{z})$.[8]

Furthermore, it seems plausible that the three test results taken on the same subject matter may still be dependent even after conditioning on U; but (3) implies that knowing underlying schooling ability would make achieved test score results, say at 7 in math, useless for predicting test score results in math at 11 or 16, and the same applies to the conditional joint distribution of the reading tests. This imposes a fairly strong duty on U.[9] We then weaken

[8] Huang and Bandeen-Roche [27] explain how a finite mixture model can be identified and estimated in the presence of continuous and discrete covariates, under the local independence assumption.

[9] A similar extension has been used, for example, by Stanghellini and van der Heijden [34].

(3) by considering dependence on covariates and allowing some conditional dependencies:[10]

Assumption 1.

$$\Pr(\boldsymbol{y} \mid \boldsymbol{b}) = \sum\nolimits_u \Pr(u \mid \boldsymbol{b}) \Pr(ol \mid u, \boldsymbol{x}) \Pr(\boldsymbol{r} \mid u) \Pr(\boldsymbol{m} \mid u), \tag{4}$$

where $\boldsymbol{m} = (m_{16}, m_{11}, m_7)$ *and* $\boldsymbol{r} = (r_{16}, r_{11}, r_7)$*.*

Assumption 1 makes it explicit the way in which family background influences observed responses:

- Writing $\Pr(\boldsymbol{r}|u)\Pr(\boldsymbol{m}|u)$ as independent of $\boldsymbol{b}$ implies that we are using a kind of *absolute performance* in math and reading ability without correcting for different backgrounds to identify U; moreover, while in the spirit of latent class analysis math and reading test scores are still used independently, residual association is allowed within $\boldsymbol{m}$ and $\boldsymbol{r}$.
- On the other hand we allow the marginal distribution of U to depend on family background variables. The intended interpretation of this individual heterogeneity is as "*potential for schooling performance*".
- Since U depends also on parents endowments U^f and U^m which are not observed, the estimates of the effect of $\boldsymbol{b}$ on U are likely to be biased. However, the omission of U^f and U^m does not bias the effect of $\boldsymbol{b}$ on OL given U (see Appendix section 6.1), which is the focus of the paper.
- We model the conditional distribution $\Pr(OL \mid U = u, \boldsymbol{x})$ to estimate the effect of a change in parents' schooling $\boldsymbol{x}$ on OL while controlling for U.

As it can be seen from equation (4), the distribution of $\boldsymbol{Y}$ conditional on U and $\boldsymbol{b}$ is decomposed into three conditionally independent blocks, namely OL results and math and reading test scores. Our second assumption is that math and reading test scores follow a first order Markov recursive system in the spirit of Griliches [22] and Griliches and Mason [23]:

Assumption 2.

$$\Pr(\boldsymbol{m} \mid u) = \Pr(m_7 \mid u) \Pr(m_{11} \mid m_7, u) \Pr(m_{16} \mid m_{11}, u);$$
$$\Pr(\boldsymbol{r} \mid u) = \Pr(r_7 \mid u) \Pr(r_{11} \mid r_7, u) \Pr(r_{16} \mid r_{11}, u).$$

[10] A latent class model where both the responses and the latent variable are allowed to depend on covariates, and residual association is allowed on responses, is described in Bartolucci and Forcina [4] in the context of capture/recapture models.

Let now $\boldsymbol{p}(\boldsymbol{b}) \in \Delta_{(m+1)\cdot 2^7}$ denote the probability vector which describes the conditional distribution of $(U, \boldsymbol{Y} \mid \boldsymbol{b})$; and let

$$\Delta^o(\boldsymbol{b}) = \{\boldsymbol{p}(\boldsymbol{b}) \in \Delta_{(m+1)\cdot 2^7} \mid \boldsymbol{p}(\boldsymbol{b}) \text{ satisfies Assumptions (1)–(2)}\}$$

denote the set of $\boldsymbol{p}(\boldsymbol{b})$'s which can be decomposed according to Assumptions 1 and 2. The dimension of this set is equal to $m + (m+1)\cdot(7+2+2)$ because there are m marginal weights for U and the conditional distribution of $Y \mid U = u$ has 11 (instead of 7) logits since $m_{11}, m_{16}, r_{11}, r_{16}$ require 2 logits each because of the Markov property; we let $v = m + (m+1)\cdot 11$.

Finally, let $\boldsymbol{\lambda}(\boldsymbol{b})$ denote the v-dimensional vector that collects the following logits and global logits:

$$\begin{aligned}\boldsymbol{\lambda}(\boldsymbol{b}) \;=\;& [\gamma^1_{U|\boldsymbol{b}} \dots \gamma^m_{U|\boldsymbol{b}}, \lambda_{M_7|0} \dots \lambda_{M_7|m}, \lambda_{R_7|0} \dots \lambda_{R_7|m}, \lambda_{M_{11}|M_7,0} \dots \lambda_{M_{11}|M_7,m},\\ & \lambda_{R_{11}|R_7,0} \dots \lambda_{R_{11}|R_7,m}, \lambda_{M_{16}|M_{11},0} \dots \lambda_{M_{16}|M_{11},m},\\ & \lambda_{R_{16}|R_{11},0} \dots \lambda_{R_{16}|R_{11},m}, \lambda_{OL|0,\boldsymbol{x}} \dots \lambda_{OL|m,\boldsymbol{x}}]'.\end{aligned}$$

Using recent advances from the theory of marginal modeling, it can be shown that any conditional distribution $\boldsymbol{p}(\boldsymbol{b}) \in \Delta^o(\boldsymbol{b})$ can be conveniently parameterized in terms of $\boldsymbol{\lambda}(\boldsymbol{b})$ without imposing any parametric restrictions besides those implied by the Assumptions (1)–(2):

Proposition 1. *The mapping from $\boldsymbol{p}(\boldsymbol{b})$ to $\boldsymbol{\lambda}(\boldsymbol{b})$ is invertible and differentiable for any $\boldsymbol{p}(\boldsymbol{b}) \in \Delta^o(\boldsymbol{b})$ with strictly positive elements.*

Proof. The result is a special case of Theorem 1 in Bartolucci–Colombi–Forcina [3] who study the properties of a general class of marginal parameterizations which constitute *link functions*, that is re-parameterizations which are one to one and at least twice differentiable. □

The mapping from $\boldsymbol{p}(\boldsymbol{b})$ to $\boldsymbol{\lambda}(\boldsymbol{b})$ can be written in explicit form by constructing an appropriate contrast matrix $\boldsymbol{C}$ (whose rows have elements summing to zero) and a marginalization matrix $\boldsymbol{M}$ (a matrix made of 1's and 0's) such that, for any $\boldsymbol{p}(\boldsymbol{b}) \in \Delta^o(\boldsymbol{b})$, we have

$$\boldsymbol{\lambda}(\boldsymbol{b}) = \boldsymbol{C} \log(\boldsymbol{M}\boldsymbol{p}(\boldsymbol{b})).$$

In particular, $\boldsymbol{C}$ is a block diagonal matrix with elements equal to $(-1 \quad 1)$. For each block of $\boldsymbol{C}$ the matrix $\boldsymbol{M}$ has two rows of length $(m+1)\cdot 2^7$ which select the elements of $\boldsymbol{p}(\boldsymbol{b})$ that constitute the two events to be compared in the corresponding logit. A Matlab routine which constructs such matrices is available from the authors.

Proposition 1 implies that the mapping from $\boldsymbol{p}(\boldsymbol{b}) \in \Delta^o(\boldsymbol{b})$ to $\boldsymbol{\lambda}(\boldsymbol{b})$ defines an invertible link function $\boldsymbol{h} : \mathbb{R}^v \mapsto \Delta^o(\boldsymbol{b})$ such that any $\boldsymbol{p}(\boldsymbol{b}) \in \Delta^o(\boldsymbol{b})$ can be written as $\boldsymbol{p}(\boldsymbol{b}) = \boldsymbol{h}\big(\boldsymbol{\lambda}(\boldsymbol{b})\big)$. Thus, for each $\boldsymbol{b}$, $\boldsymbol{\lambda}(\boldsymbol{b})$ conveys all relevant information on $\boldsymbol{p}(\boldsymbol{b})$. When the covariates are discrete with a very limited number of distinct configurations (strata), we could define a finite number of dummy variables for mutually exclusive and exhaustive configurations, and use these in place of $\boldsymbol{b}$. In practice this would be equivalent to fit a separate model to each stratum and the model may be called *saturated* (see e.g. Wooldridge [36] p.456 for terminology). Since in this case there would be no sharing of parameters across strata, this is equivalent to imposing no restriction on how $\boldsymbol{\lambda}(\boldsymbol{b})$ depends on $\boldsymbol{b}$. The advantage is that there is no problem of mispecification in the mapping from $\boldsymbol{b}$ to $\boldsymbol{p}(\boldsymbol{b})$; in other words, under Assumptions 1–2, the saturated model is non parametric.

Typically however covariates may be continuous or take on so many values that most strata contain only one subject, so that the approach just described is not viable. We model $\boldsymbol{\lambda}(\boldsymbol{b})$ as a linear function of the covariates, hence we estimate the following multivariate regression system:

$$\begin{aligned}
&\Pr(OL = 1 \mid u, \boldsymbol{x}) = \Lambda\big(a_{OL}(u) + \boldsymbol{x}'\boldsymbol{\beta}_{OL}\big)\\
&\Pr(M_{16} = 1 \mid M_{11}, u) = \Lambda\big(a_{M_{16}}(u) + b_{M_{16}}(u)M_{11}\big)\\
&\Pr(R_{16} = 1 \mid R_{11}, u) = \Lambda\big(a_{R_{16}}(u) + b_{R_{16}}(u)R_{11}\big)\\
&\Pr(M_{11} = 1 \mid M_7, u) = \Lambda\big(a_{M_{11}}(u) + b_{M_{11}}(u)M_7\big)\\
&\Pr(R_{11} = 1 \mid R_7, u) = \Lambda\big(a_{R_{11}}(u) + b_{R_{11}}(u)R_7\big)\\
&\Pr(M_7 = 1 \mid u) = \Lambda\big(a_{M_7}(u)\big)\\
&\Pr(R_7 = 1 \mid u) = \Lambda\big(a_{R_7}(u)\big)\\
&\Pr(U \geq u \mid \boldsymbol{b}) = \Lambda\big(a_U(u) + \boldsymbol{b}'\boldsymbol{\beta}_U\big)\,,
\end{aligned} \tag{5}$$

where $\Lambda(t) = e^t/(1 + e^t)$ denotes the logit link function. This can be written more compactly as

$$\boldsymbol{\lambda}(\boldsymbol{b}) = \boldsymbol{B}\boldsymbol{\psi} \tag{6}$$

where $\boldsymbol{B}$ is a design matrix whose dependence on $\boldsymbol{b}$ reflects the effect of the covariates on the different elements of the joint distribution, and $\boldsymbol{\psi}$ is the vector of model parameters.

The standard approach to parameters' estimation in finite mixture models is the *EM* algorithm whose implementation to our context is described in the

Appendix. The basic idea of the EM algorithm is to consider that, if the joint frequency table $(U, \boldsymbol{Y})$ were known, maximum likelihood estimation would be equivalent to estimation of a regression model within the multinomial distribution. At the E (expectation) step the unobservable frequency table is replaced by the expected value computed conditionally on the observed frequency table and the (possibly updated) estimates of the model parameters. It has been shown (Dempster–Laird–Rubin [18]) that the algorithm converges to the maximum of the true likelihood. Details on the identification and estimation of the model parameters $\boldsymbol{\psi}$ are contained in the Appendix.

4. Results

4.1. **Parameters' estimation.** We start by estimating model (6) under different numbers of latent classes. Maximum likelihood estimation is performed by a EM algorithm as described in the Appendix; we have written a Matlab program which implements it, available upon request. The maximized log-likelihood $L(\hat{\boldsymbol{\psi}})$ and Schwartz's Bayesian Information Criterion $BIC(\hat{\boldsymbol{\psi}}) = -2L(\hat{\boldsymbol{\psi}}) + \log(n)v$, where n denotes sample size and v is the number of parameters (which depends on the number of latent classes $m+1$), are given in Table 2 below. Since $BIC(\hat{\boldsymbol{\psi}})$ is lower with three latent classes in both samples, these results seem to indicate that three latent classes are adequate to represent unobserved heterogeneity. A comparison of the estimated coefficients in the OL equation, which are of major interest for our purposes, reveals that while with two latent classes the coefficients for parents' schooling are sensibly greater than those with three classes, there is practically no difference between the estimated coefficients for parents' schooling with three or four classes.[11]

		Daughters		Sons	
latent cl.	param.	$L(\hat{\boldsymbol{\psi}})$	$BIC(\hat{\boldsymbol{\psi}})$	$L(\hat{\boldsymbol{\psi}})$	$BIC(\hat{\boldsymbol{\psi}})$
2	30	-8014.99	16261.47	-8293.13	16818.58
3	42	-7875.06	16074.21	-8166.58	16658.41
4	54	-7845.17	16107.03	-8121.37	16660.92

Table 2. Maximized log-likelihood and BIC

Next, for parsimony and sharpness of parameters' estimation, we impose the restriction that the slopes $b_{M11}(u), b_{R11}(u), b_{M7}(u), b_{R7}(u)$, which give the

[11] Estimated coefficients for two and four latent classes are not reported for the sake of brevity.

recursive structure to test results, do not depend on U. Since U takes three levels, this imposes a total of $4 \times 2 = 8$ linear constraints. The restriction is not rejected by a standard LR test (p-values are respectively equal to 0.6875 and 0.2764 in the males and females subsamples).

Table 3 below contains our estimates.[12] We observe that:

- The coefficients of the parents' education variables in the OL equation imply the presence of the role effects as described in the introduction: after controlling for the child's schooling endowments, the father's education significantly helps his son's chance of achieving OL certification, and not his daughter's; on the other hand, mothers' education has no statistically significant effects.
- There is a strong positive association between child's endowment U and family background characteristics. However, as we discuss in the Appendix, since we do not control for parents' endowments, our estimate of $\boldsymbol{\beta}_U$ may be biased, and the causes for high U are left unexplored. Nevertheless, the reasons why this bias should be significantly different for the two parents are unclear.
- In general, the effect of having a high rather than low level of U is rather substantial on all response variables; to appreciate its quantitative impact, notice that, in the logit scale, a change of value say from -2 to +2 implies a change in the probability of success from about 12% to 88%, while a change from -1 to +1 implies a change from 27% to 73%.
- Even after conditioning on U, there is still a strong positive correlation between test score results in the same subject taken at different ages, as it emerges from the significantly positive estimates of $b_{M11}, b_{R11}, b_{M7}, b_{R7}$.

Finally, for the sake of comparison, we have estimated the previous model on the pooled sample of sons and daughters. As can be seen by comparing the corresponding lines of Tables 3 and 4 below, the pooled sample estimates are approximately equal to the average of the corresponding estimates for daughters and sons, suggesting a somewhat stronger role of fathers. We have shown that considering sons and daughters subsamples separately is crucial for clarifying the nature of this asymmetry.

[12] To ease interpretation of the intercepts, all covariates have been centered.

	Daughters		Sons	
	coeff	se	coeff	se
$\boldsymbol{\beta}_{OL}$				
fs	-0.0119	0.0519	0.1461	0.0526
ms	0.0771	0.0597	0.0305	0.0595
$\boldsymbol{\beta}_{U}$				
fs	0.2635	0.0410	0.2995	0.0415
ms	0.2241	0.0454	0.2652	0.0462
fi	1.0677	0.1496	1.2677	0.1539
mi	0.4859	0.1601	0.028	0.1578
ty	0.0435	0.0083	0.028	0.0085
Other parameters				
$a_{OL}(U=0)$	-2.5911	0.2323	-2.5651	0.2273
$a_{OL}(U=1)$	0.3507	0.1094	-0.2091	0.1113
$a_{OL}(U=2)$	2.4298	0.1920	2.1521	0.1860
$a_{M_{16}}(U=0)$	-2.9577	0.2298	-2.3952	0.2121
$a_{M_{16}}(U=1)$	-0.7126	0.1464	0.1247	0.1505
$a_{M_{16}}(U=2)$	3.1312	0.5464	4.9600	1.4623
$b_{M_{11}}$	0.5764	0.1911	0.4196	0.1851
$a_{R_{16}}(U=0)$	-2.9529	0.2310	-2.2093	0.1608
$a_{R_{16}}(U=1)$	-0.5772	0.1372	-0.5355	0.1346
$a_{R_{16}}(U=2)$	1.7708	0.3043	1.8477	0.3594
$b_{R_{11}}$	1.4707	0.1605	1.9963	0.1605
$a_{M_{11}}(U=0)$	-3.2631	0.3123	-3.3529	0.3536
$a_{M_{11}}(U=1)$	-0.0242	0.1323	-0.2521	0.1421
$a_{M_{11}}(U=2)$	3.4093	0.4224	3.2757	0.4574
b_{M_7}	0.3324	0.1579	0.6199	0.1528
$a_{R_{11}}(U=0)$	-3.0733	0.2504	-2.9243	0.2784
$a_{R_{11}}(U=1)$	-0.2783	0.1485	-0.1541	0.1350
$a_{R_{11}}(U=2)$	1.8328	0.2227	2.0751	0.2319
b_{R_7}	0.6454	0.1499	0.5282	0.1484
$a_{M_7}(U=0)$	-1.8338	0.1313	-1.4935	0.1200
$a_{M_7}(U=1)$	-0.3027	0.0961	-0.0856	0.0973
$a_{M_7}(U=2)$	0.9859	0.0941	1.2074	0.0987
$a_{R_7}(U=0)$	-1.5506	0.1267	-2.2298	0.1760
$a_{R_7}(U=1)$	0.6251	0.1052	-0.1301	0.1067
$a_{R_7}(U=2)$	2.0281	0.1340	1.7173	0.1263
$a_U(U=1)$	1.0530	0.0816	1.0468	0.0941
$a_U(U=2)$	-0.7378	0.0907	-0.7335	0.0887

TABLE 3. Parameters' estimates

	coeff	se
fs	0.0704	0.0368
ms	0.0589	0.0417

Table 4. $\boldsymbol{\beta}_{OL}$ estimates in the pooled sample of sons and daughters.

4.2. **Direct effects.** We now translate the above estimates into a quantitative appraisal of the role effect. We consider the effect on OL attainment of increasing a parent education by three full years of schooling, leaving unchanged the schooling of the other parent, starting from a situation where both parents have an average level of schooling.

The average effect of increasing a parent's education for a given level of child's endowment U can be calculated as:

$$\begin{aligned}\delta(u)_{S^i} &= Pr(OL = 1 \mid S^j = \mu^j, S^i = \mu^i + 3, U = u) - \\ &\quad Pr(OL = 1 \mid S^j = \mu^j, S^i = \mu^i, U = u) \\ &= \Lambda(a_{OL}(u) + 3\beta_{OL,S^i}) - \Lambda(a_{OL}(u)), \quad i, j = f, m, \quad u = 0, 1, 2,\end{aligned}$$

where the second equation follows since we have centered father's and mother's schooling. These effects can be consistently estimated using the coefficients in the OL-equation; furthermore, using the estimated variance matrix of $(a_{OL}(u), \beta_{OL,S^i})$, an asymptotic standard error can be derived by application of the delta method. Their numerical values are in the following table:

	Daughters		Sons	
	$\delta(u)_{S^f}$	se	$\delta(u)_{S^f}$	se
U=0	-0.0023	0.0098	0.0351	0.0169
U=1	-0.0087	0.0380	0.1091	0.0392
U=2	-0.0027	0.0137	0.0344	0.0122
	$\delta(u)_{S^m}$	se	$\delta(u)_{S^m}$	se
U=0	0.0166	0.0148	0.0063	0.0129
U=1	0.0547	0.0413	0.0227	0.0442
U=2	0.0156	0.0116	0.0082	0.0156

Table 5. Direct Effects of Father's and Mothers' Schooling

The table shows that the direct effect, measured as difference in probabilities of achievement, is rather substantial and statistically significant only for fathers on sons. Also interesting to notice is the fact that direct effects are highly nonlinear in U.

4.3. **A quasi-saturated model.** Since some of the covariates in $\boldsymbol{b}$ are either discrete and take on many values or are actually continuous, fitting a saturated model would be impossible; to check the robustness of our results we now present a model which may be seen as a feasible approximation of the saturated model. We dichotomize parents' schooling into two dummy variables *fsh* and msh, which take value 1 if the parent has left school after 16 years of age; family income is also dichotomized into a dummy variable tyh which takes value 1 when family income is above the sample median. The set of covariates can then be arranged into a $2^5 = 32$ different covariate configurations so that a saturated model could, in principle, be fitted. These 32 dummy variables can be arranged into a null effect, 5 main effects $(\mathit{fsh}, msh, \mathit{fi}, mi, tyh)$, 10 second-order interactions, 10 third-order interactions, 5 fourth-order interactions and 1 fifth order interaction. Because in both samples there was very high collinearity between interactions of order higher than the second, we fitted a *quasi-saturated model* by ignoring all interactions beyond the second.

We thus re-estimate model (5) where $\boldsymbol{b}$ in the U-equation contains 15 dummy variables, while $\boldsymbol{x}$ in the OL-equation contains the three dummies $(\mathit{fsh}, msh, \mathit{fsh}\cdot msh)$. We again use the EM algorithm to estimate this model, using the same parameters' complexity as the model just estimated. The value of the maximized log-likelihood for the females and males samples is equal to -7892.05 and -8190.56 respectively; thus even though there is a higher number of slope parameters, the quasi-saturated model has a worse fit than the previous model, the reason being that whatever is gained by the increased number of parameters is counterbalanced by the loss of information due to the dichotomization of covariates. The estimated coefficients for the effects of parents' schooling on OL are reported in the table below:[13]

	Daughters		Sons	
	coeff	se	coeff	se
fs	-0.1450	0.2719	0.5600	0.2869
ms	-0.0678	0.3001	0.0908	0.2857
fs · ms	0.5591	0.5016	0.8691	0.5872

Table 6. $\boldsymbol{\beta}_{OL}$ estimates in the quasi-saturated model

A comparison of these estimated coefficients with those of Table 3 above reveals that the main message of the model discussed in the previous section

[13] The other estimated parameters are similar to the previous model and are not reported for the sake of brevity.

is quite robust. Conditionally on unobservable child's endowment, father's schooling has still a positive effect on the probability that the child's attains the O-Level certification; while all other coefficients are not statistically significant, it may be interesting to notice some positive (but insignificant) effect in the interaction term.

5. Concluding Remarks

Recent papers [14] have analyzed the causal effect of parents' education on children's controlling for parents' unobserved endowment, by use of twin parents, adoptees or compulsory schooling law instruments. By applying recently developed finite mixture models to the UK NCDS dataset, we control for the child's own schooling endowments, by exploiting information on early cognitive tests.

By conditioning on child's ability rather than on parents' we measure the direct effect of parents' schooling on children's education. The effect of parents' schooling on children's education attainment *given* the latter's ability reflects parental pressure, and can thus be interpreted as a 'role' effect. To allow for the possibility of its dependence on gender we consider sons and daughters subsamples separately. We find that only fathers' education matters, but that its impact is entirely confined to the education of their sons.

This result may well reflect the social structure of Western families in the seventies (the data domain), and if the women's role has changed a different picture may emerge from more recent data. But the message we get from our findings remains that children can respond strongly to family pressure on schooling attainment. From a policy –or rather 'cultural'– viewpoint this suggests that when parents' pressure is weak only the social environment, school primarily, can make up for this loss by helping the young to appreciate the value of education.

6. Appendix

6.1. **Marginal causal effect.** Consider the marginal causal effect of an ordered discrete variable X on a binary variable Y

$$\Delta_x \equiv \Pr(Y = 1|X = x + 1) - \Pr(Y = 1|X = x),$$

[14] See for example [1], [7], [8], [10], [11], [15], [17], [29], [31], [32], [33].

where we have omitted other covariates for simplicity. In terms of the present paper, Y represents scholastic attainment and X parent's years of schooling. Assume also the existence of two discrete variables U and V, which are meant to capture respectively child's and parent's unobservable endowments.

Δ_x above may be expanded by considering that $\Pr(Y|X) = \sum_u \Pr(Y, U = u|X)$; letting $P_{x,u} = \Pr(Y = 1|X = x, U = u)$ and $Q_{u|x} = \Pr(U = u|X = x)$, one obtains $\Delta_x = \sum_u (P_{x+1,u} Q_{u|x+1} - P_{x,u} Q_{u|x})$. By adding and subtracting $\sum_u P_{x,u} Q_{u|x+1}$ to this expression, rearranging terms and noting that if U has $m+1$ levels we may write $Q_{0|x} = 1 - \sum_1^m Q_{u|x}$, one gets

$$\Delta_x = \sum\nolimits_{u=0}^{m} (P_{x+1,u} - P_{x,u}) Q_{u|x+1} + \sum\nolimits_{u=1}^{m} (P_{x,u} - P_{x,0})(Q_{u|x+1} - Q_{u|x}).$$

The first component is a weighted average of the direct effect of X on Y. The second term is the sum of the products of the effect of X on U times the effect of U on Y, which may be interpreted as a measure of the indirect effect of X on Y carried to Y through U. Though the above decomposition is not unique since, for instance, in the first component $Q_{u|x}$ could be used as weights with a minor change in the second component, it may be considered as a discrete analogue of the decomposition $\beta = b + ce$ in the linear system (1)–(2).

If subjects could be classified according to their value of U, the direct causal effect of X on Y could alternatively be evaluated at the various values of U as $P_{x+1,u} - P_{x,u}$ which we denote by

$$\delta_x(u) \equiv \Pr(Y = 1|X = x + 1, U = u) - \Pr(Y = 1|X = x, U = u);$$

if U had only few levels, inspection of the individual values of $\delta_x(u)$ would be more instructive than looking at their average. This can be compared to the simple linear model $E(Y \mid X, U) = a + bX + cU$, where clearly the single coefficient b captures the direct effect.

Now consider the additional unobservable variable V, and assume it is *ignorable*:

$$\Pr(Y = 1 \mid X = x, U = u, V = v) = \Pr(Y = 1 \mid X = x, U = u),$$

so that Y is independent of V given U, X. Then a similar argument to the one used above may be applied to expand the effect of X on U in the second component. To do so it is convenient to write $Q_{u|xv} = \Pr(U = u \mid X = x, V =$

v) and $R_{v,x} = \Pr(V = v \mid X = x)$; assuming that V has $h+1$ levels one obtains

$$\begin{aligned}\Delta_x = &\sum\nolimits_{u=0}^{m}(P_{x+1,u} - P_{x,u})Q_{u|x+1}\\ &+\sum\nolimits_{u=1}^{m}(P_{x,u} - P_{x,0})\sum\nolimits_{v=0}^{h}(Q_{u|x+1,v} - Q_{u|x,v})R_{v|x+1}\\ &+\sum\nolimits_{u=1}^{m}(P_{x,u} - P_{x,0})\sum\nolimits_{v=1}^{h}(Q_{u|x,v} - Q_{u|x,0})(R_{v|x+1} - R_{v|x}),\end{aligned}$$

which makes it clear that, if X, V are independent, the third component is 0 so that the marginal effect can be decomposed into a direct and indirect effect of X on Y. If instead there is correlation within both X, V and U, V, then the third component is nonzero, and thus knowledge of U is not sufficient for decomposing the marginal effect into a direct and indirect effect.

Summing up, if U is known but V is not, under ignorability of V one can estimate the direct causal effect of X on Y given U, but one cannot obtain an unbiased estimate of the marginal causal effect of X on Y and its decomposition into a direct and indirect component, since the term $P_{x+1,u} - P_{x,u}$ requires controlling for U, while the other terms in Δ_x require control of V.

6.2. **Identification, estimation and computation of model** (6)**.**

6.2.1. *Likelihood inference.* **The true log-likelihood.** Let $\boldsymbol{n}(i)$ be the 2^7 vector containing the frequency table of the response variables $\boldsymbol{Y}$ in lexicographic order for the subjects with covariate $\boldsymbol{b}_i$; if there is a single subject with such features, $\boldsymbol{n}(i)$ is a vector of 0s except for a 1 in the cell corresponding to the response pattern $\boldsymbol{y}(i)$. Let also $\boldsymbol{q}(i)$ denote the probability distribution for subjects with covariate $\boldsymbol{b}_i$. The log-likelihood may be written as

$$L = \sum L_i = \sum \boldsymbol{n}(i)' \log[\boldsymbol{q}(i)].$$

The latent log-likelihood. Let $\boldsymbol{L} = (\boldsymbol{1}_{m+1}{}' \otimes \boldsymbol{I}_{2^7})$ denote the matrix which marginalizes with respect to the latent variable U, $\boldsymbol{p}(i)$ the vector containing the joint probability distribution of $(U, \boldsymbol{Y})$ for subjects with covariate $\boldsymbol{b}_i$ and $\boldsymbol{m}(i)$ the vector containing the unobservable frequency table of $(U, \boldsymbol{Y})$ for subjects with covariate $\boldsymbol{b}_i$. Clearly $\boldsymbol{n}(i) = \boldsymbol{L}\boldsymbol{m}(i)$ and $\boldsymbol{q}(i) = \boldsymbol{L}\boldsymbol{p}(i)$.

If the latent class U could be observed, the corresponding log-likelihood would have the form

$$\Lambda = \sum \Lambda_i = \sum \boldsymbol{m}(i)' \log[\boldsymbol{p}(i)].$$

Maximizing this log-likelihood is as a problem of incomplete data which may be tackled by the EM algorithm (Dempster–Laird–Rubin [18]).

The E step. Because the multinomial is a member of the exponential family, the conditional expectation involved in the E step is equivalent to computing the so called posterior probability of latent class U given the observed configuration $\boldsymbol{y}$

$$\Pr(U \mid \boldsymbol{y}, \boldsymbol{z}_i) = \frac{p(i, U, \boldsymbol{y})}{q(i, \boldsymbol{y})}$$

so that $m(i, u, \boldsymbol{y}) = n(i, \boldsymbol{y}) \Pr(u \mid \boldsymbol{y}, \boldsymbol{b}_i)$ follows from a simple expectation of a multinomial distribution for U.

The M step. Implementation of the method of scoring for the maximization of Λ with respect to the model parameters $\boldsymbol{\psi}$ requires computation of the score vector (first derivative with respect to $\boldsymbol{\psi}$) and of the expected information matrix (minus the expected value of the second derivative). Since Λ is a multinomial log-likelihood, exponential family results can be exploited to make such calculations straightforward. In practice, after rewriting Λ in terms of the canonical parameters of the multinomial distribution, say $\boldsymbol{\theta}(i)$, there are invertible and differentiable mappings from $\boldsymbol{\theta}(i)$ to the vector of probabilities $\boldsymbol{p}(i)$ and from $\boldsymbol{p}(i)$ to $\boldsymbol{\lambda}(i)$ (the latter mapping is described after Proposition 1), while $\boldsymbol{\lambda}(i)$ is linked to $\boldsymbol{\psi}$ by the linear regression model. The interested reader may see Dardanoni and Forcina [16] or Bartolucci–Colombi–Forcina [3] for details.

6.2.2. *Computational issues.* The EM algorithm is a very robust method of estimation of the model parameters for latent class models. However, it suffers from at least two drawbacks: it can be very slow with large data sets, and, by itself, does not provide a consistent estimate of the variance-covariance matrix of the model parameters. This is so because the expected information matrix of the latent likelihood is based on the assumption that the vector $\boldsymbol{m}(i)$ is known, using its inverse as an estimate of the variance matrix implies that standard errors will in general be underestimated. The correct information matrix may be computed by differentiating the incomplete data likelihood as follows. Write $L_i = \boldsymbol{n}(i)'\tilde{\boldsymbol{G}}\boldsymbol{\gamma}_i - n_i \log[\boldsymbol{1}' \exp(\tilde{\boldsymbol{G}}\boldsymbol{\gamma}_i)]$ where $\boldsymbol{\gamma}_i$, the canonical parameter of the observed multinomial, may be written as $\tilde{\boldsymbol{H}} \log[\boldsymbol{L} \exp(\boldsymbol{G}\boldsymbol{\theta}_i)/\boldsymbol{1}' \exp(\boldsymbol{G}\boldsymbol{\theta}_i)]$, where $\tilde{\boldsymbol{H}}$ is a $t \times (t-1)$ contrast matrix used to define the canonical parameters and $\tilde{\boldsymbol{G}}$ is its right inverse while $\boldsymbol{G}$ is the design matrix which defines the canonical parameters $\boldsymbol{\theta}$ for the latent distribution $\boldsymbol{p}(i)$ which has v columns of full rank. By differentiating L_i by the chain rule with respect to $\boldsymbol{\psi}$ one may write

$$\sum \frac{\partial L_i}{\partial \boldsymbol{\psi}} = \boldsymbol{B}_i{}' \boldsymbol{R}_i{}' \boldsymbol{G}' \boldsymbol{\Omega}_i \boldsymbol{L}' diag(\boldsymbol{q}_i)^{-1} \tilde{\boldsymbol{H}}' \tilde{\boldsymbol{G}}' (\boldsymbol{n}(i) - n_i \boldsymbol{q}_i)$$

where $n_i = \boldsymbol{1}' \boldsymbol{n}(i)$, $\boldsymbol{\Omega}_i = diag[\boldsymbol{p}(i) - \boldsymbol{p}(i)\boldsymbol{p}(i)']$ and $\boldsymbol{R}_i$ is the derivative of the canonical parameter $\boldsymbol{\theta}_i$ with respect to $\boldsymbol{\lambda}_i{}'$. Because $E(\boldsymbol{n}(i) - n_i \boldsymbol{q}_i) = \boldsymbol{0}$, minus the expected value of the second derivative may be written as

$$\boldsymbol{F}_i = \boldsymbol{B}_i{}' \boldsymbol{R}_i{}' \boldsymbol{G}' \boldsymbol{\Omega}_i \boldsymbol{L}' diag(\boldsymbol{q}_i)^{-1} \tilde{\boldsymbol{H}}' \tilde{\boldsymbol{G}}' \boldsymbol{L} \boldsymbol{\Omega}_i \boldsymbol{R}_i \boldsymbol{B}_i$$

(where $\tilde{\boldsymbol{H}}' \tilde{\boldsymbol{G}}'$ is simply equal to $\boldsymbol{I}_t - \boldsymbol{1}_t \boldsymbol{1}_t{}'/t$ with $t = 2^7$), so the information matrix is simply $\sum_i \boldsymbol{F}_i$.

6.3. Model identifiability.

A statistical model depending on a vector of parameters $\boldsymbol{\psi} \in \Psi$ is identifiable if there is no subset of Ψ where the likelihood is constant. When, as in our case, the likelihood is differentiable, the model is identifiable if there is no $\psi_0 \in \Psi$ such that the matrix of second derivatives (or equivalently the observed information matrix) computed at $\boldsymbol{\psi}_0$ is not of full rank. To analyze the identifiability of our model, consider first the parametrization in terms of the vector $\boldsymbol{\gamma}$ obtained by stacking the vectors $\boldsymbol{\gamma}_i$ of canonical parameters of the saturated log-linear model for each subject in the manifest distribution. Since this saturated log-linear model is identifiable, the problem consists in checking that lack of identifiability is not introduced by the different parametric transformations:

(1) from $\boldsymbol{\gamma}$ to $\boldsymbol{\theta}$, the vector obtained by stacking the vectors of canonical parameters of the latent class model for each subject,
(2) from $\boldsymbol{\theta}$ to $\boldsymbol{\lambda}$, the vector obtained by stacking the vectors of marginal parameters for each subject,
(3) the regression model $\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\psi}$.

The identifiability of the regression model is easily established by checking that $\boldsymbol{B}$ is of full column rank. Results from Bartolucci, Colombi and Forcina (2006) ensure that the transformation to the marginal parameters is invertible and differentiable. So, the crucial transformation is the first one and, because the transformation is at the subject's level, to prove identifiability it is sufficient to show that the following matrix of first derivatives is of full column rank

$$\boldsymbol{T}_i = \frac{\partial \gamma_i}{\partial \boldsymbol{\theta}_i{}'} = \tilde{\boldsymbol{H}} diag(\boldsymbol{q}_i)^{-1} \boldsymbol{L} \boldsymbol{\Omega} \boldsymbol{G}.$$

In most cases, $\boldsymbol{T}_i$ will be a rectangular matrix of full column rank, which implies that $\boldsymbol{\gamma}_i$ is constrained within a non linear subspace determined by the specific latent class model. Unfortunately there is no general result in the latent class literature to establish whether a latent class model is identifiable, though identifiability of most models of interest under local independence is known. When, like for the model used in this paper, results are not available, the matrix $\boldsymbol{T}_i$ above may be easily computed for a randomly chosen set of values of $\boldsymbol{\theta}_i$ and the full rank condition checked. This is what we have done for our model by sampling a reasonable number of $\boldsymbol{\theta}_i$'s from a standard normal. As we could not find a single instance where the rank was any close to being deficient, we have good practical evidence to believe that the model is indeed identifiable even for a single subject. However, notice that even when the matrix $\boldsymbol{T}_i$ is not of full column rank, the model may become identifiable due to the regression component. The reason for this is that the regression model implies that most parameters are shared in the probability distribution of different subjects, so that information may be collected from different subjects.

References

[1] Antonovics, Kate L. and Arthur S. Goldberger (2005): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Comment", *American Economic Review* **95**, pp. 1738–1744

[2] Arcidiacono, Peter and John Bailey Jones (2003): "Finite mixture distributions, sequential likelihood and the EM algorithm", *Econometrica* **71**, pp. 933–946

[3] Bartolucci, Francesco, Roberto Colombi and Antonio Forcina (2006): "An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints", *Statistica Sinica*, to appear

[4] Bartolucci, Francesco and Antonio Forcina (2006): "A class of latent marginal models for capture-recapture data with continuous covariates", *Journal of the American Statistical Association* **101**, pp. 786–794

[5] Bergsma, Wicher and Tamàs Rudas, (2002): "Marginal models for categorical data" *Annals of Statistics* **30**, pp. 140–159

[6] Behrman, Jere R. (1997): "Women's Schooling and Child Education: a Survey", *mimeo*

[7] Behrman, Jere R. and Mark R. Rosenzweig (2002): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation?", *American Economic Review* **92**, pp. 323–334

[8] Behrman, Jere R. and Mark R. Rosenzweig (2005): "Does Increasing Women's Schooling Raise the Schooling of the Next Generation? Reply", *American Economic Review* **95**, pp. 1745–1751

[9] Bergstrom, Theodore C. (1997): “A Survey of Theories of the Family”, in *Handbook of Population and Family Economics*, edited by Mark Rosenzweig and Oded Stark. Amsterdam, Elsevier

[10] Bjorklund, Anders, Mikael Lindahl, and Erik Plug (2006): “The Origins of Intergenerational Associations: Lessons from Swedish Adoption Data”, *Quarterly Journal of Economics* **121**, pp. 999–1028

[11] Black, Sandra E., Paul J. Devereux and Kjell G. Salvanes (2005): “Why the Apple Doesn’t Fall Far: Understanding Intergenerational Transmission of Human Capital”, *American Economic Review* **95**, pp. 437–449

[12] Blundell, Richard, Lorraine Dearden and Barbara Sianesi (2005): “Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey”, *Journal of the Royal Statistical Society A*, **168**, pp. 473–512

[13] Boudon, Raymond, (1974): *Education, Opportunity, and Social Inequality: Changing Prospects in Western Society* Wiley, New York

[14] Cameron, Stephen V., and James J. Heckman (1998): “Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males”, *Journal of Political Economy* **106**, pp. 262–333

[15] Chevalier, Arnaud (2004): “Parental Education and Child’s Education: A Natural Experiment”, *IZA discussion paper* n. 1153

[16] Dardanoni, Valentino and Antonio Forcina (2006): “Multivariate ordered regression”, *mimeo*

[17] Dearden, Lorraine S., Stephen Machin, and Howard Reed (1997): “Intergenerational Mobility in Britain”, *Economic Journal* **110**, pp. 47–64

[18] Dempster, Arthur P., Nan M. Laird and Donald B. Rubin (1977): “Maximum likelihood for incomplete data via the EM algorithm”, *Journal of the Royal Statistical Society* Series B **39**, pp. 1–22

[19] Duflo, Esther (2003): “Grandmothers and Granddaughters: Old Age Pension and Intra-household Allocation in South Africa”, *World Bank Economic Review*, **17**, pp. 1–25

[20] Erikson, Robert, John H. Goldthorpe, Michelle Jackson, Meir Yaish, and D. R. Cox (2005): “On class differentials in educational attainment”, *Proceedings of the National Academy of Science* **102**, pp. 9730–9733

[21] Goodman, Leo (1974): “Exploratory latent structure analysis using both identifiable and unidentifiable models”, *Biometrika* **61**, pp. 215–231

[22] Griliches, Zvi (1977): “Estimating the Returns to Schooling: Some Econometric Problems” *Econometrica* **45:1**, pp. 1–22

[23] Griliches, Zvi and William M. Mason (1972): ”Education, Income, and Ability,” *Journal of Political Economy* **80** Part II, pp. S74-S103

[24] Haveman, Robert and Barbara Wolfe (1995): “The Determinants of Children’s Attainments: a Review of Methods and Findings”, *Journal of Economic Literature* **33**, pp. 1829-1878

[25] Heckman, James and Burton Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, **52**, pp. 271-320

[26] Holmlund Helena, Mikael Lindahl and Erik Plug (2005): "Estimating Intergenerational Effects of Education: A Comparison of Methods", mimeo, Stockholm University

[27] Huang G. and K. Bandeen-Roche (2004): "Building an identifiable latent class model, with covariate effects on underlying and measured variables" *Psychometrika* **69**, pp. 5-32

[28] B. Lindsay, C. Clogg, and J. Grego (1991): "Semiparametric estimation of the Rasch model and related exponential response models, including a simple latent class model for item analysis", *Journal of the American Statistical Association* **86**, pp. 96-107

[29] Oreopoulus, Philip, Marianne Page, and Anne Huff Stevens (2006): "Human Capital Transfer from Parent to Child? The Intergenerational Effects of Compulsory Schooling", *Journal of Labor Economics*, forthcoming

[30] Pearl, Judea (2000): *Causality*, Cambridge University Press

[31] Plug, Erik (2004): "Estimating the effect of Mother's Schooling Using a Sample of Adoptees", *The American Economic Review* **94**, pp. 358-368

[32] Sacerdote, Bruce (2002): "The Nature and Nurture of Economic Outcomes", *American Economic Review* **92**, pp. 344-48

[33] Sacerdote, Bruce (2004): "What Happens When We Randomly Assign Children to Families", *NBER working paper* n. 10894

[34] Stanghellini, Elena and Peter G.M. van der Heijden (2004): "A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account", *Biometrics* **60**, pp. 510-516

[35] Vermunt, Jeroen K. and Jay Magidson (2003), "Latent class analysis", in *Encyclopedia of Research Methods for the Social Sciences*, Michael S. Lewis-Beck, Alan Bryman and Tim Futing Liao editors, Sage Publications, NewBury Park.

[36] Wooldridge, Jeffrey M. (2002): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, The MIT Press.

Facoltà di Economia, Università di Palermo
E-mail address: vdardano@unipa.it

Facoltà di Economia, Università di Perugia
E-mail address: forcina@stat.unipg.it

Facoltà di Economia, Università di Palermo
E-mail address: modica@unipa.it