Missing Ordinal Covariates with Informative Selection

Alfonso Miranda & Sophia Rabe-Hesketh

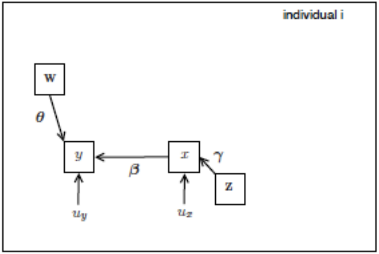
Institute of Education, University of London

Motivation

Often key ordinal explanatory variables are missing in the data for a large proportion of the sample

Mother's education is often a *"missing control"* either because no such information is available (administrative records) or because of item non-response (surveys). This missing covariate, for instance, is likely to be a *"confounder"* of the relationship between achievement and ethnic group, leading to a problem of *omitted variable bias*

Complete data case

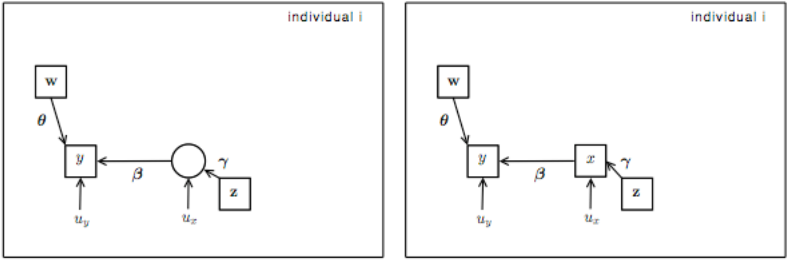


(a) Complete data

- [y]: main response variable (continuous)
- [x]: key ordinal control variable
- [$\mathbf{w}_i$]: predictors for y
- [$\mathbf{z}_i$]: predictors of x

Note: $\mathbf{w}_i = \{\mathbf{z}_i, y\text{-only explanatory variables}\}$

Missing covariate case



(b) Missing covariate

(c) Complete data

- x is missing for a large proportion of the sample!
- Selection (not deletion from sample) follows rule S

What are the consequences?

- At best: inefficient estimators
- At worst: inconsistent estimators

All depends on what is the mechanism S that cause the data to be missing!

- What can be done?

List-wise deletion (complete case analysis)

- Drop cases with $S = 0$
- Inefficient estimator!
- Consistent only if the probability of selection does not depend on y given the explanatory variables

$$\Pr(S|Y, X, V) = \Pr(S|X, V)$$

where,

$$V \equiv \text{other explanatory variables}$$

- Consistent if probability of selection depends on the missing covariate x (Grilliches 1978,Little 1992, Little and Rubin 2002, Wooldridge 2002)

$$\Pr(S|Y, X, V) = \Pr(S|X, V)$$

That is, when data is not missing at random (NMAR)

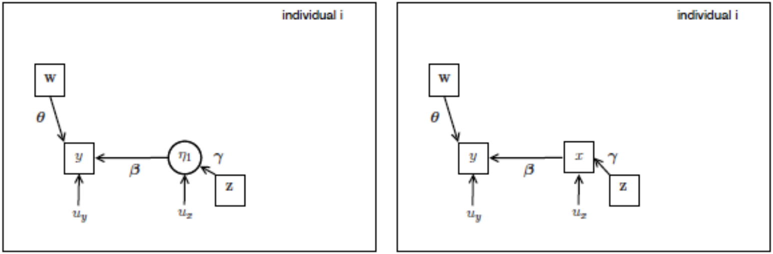
Missing at Random (MAR) — $Pr(S|Y,X,V) = P(S|Y,V)$

- Weighted complete case analysis
  - Consistent estimators if a weighted version of the estimation method is used with weights given by $P(S|Y,V)^{-1}$ (consistent even if S depends on X, i.e. when data is really NMAR)
- Multiple imputation
  - Missing data filled by sampling from the estimated regression model $\widehat{\Pr}(X|Y,V)$. Do this multiple times, yielding several imputed datasets. Each analysed by conventional methods and estimates averaged across datasets.
  - More efficient than complete case analysis but inconsistent if selection depends on X (i.e., when data is in fact NMAR)
- Maximum likelihood
  - Joint model for Y and X, with missing values of X integrated out (EM algorithm)
- Bayesian estimation of joint model for Y and X (sample X from its posterior distribution along other parameters)

Not Missing at Random (NMAR) — $Pr(S|Y, X, V)$

- Any of the methods suitable for MAR deliver inconsistent estimators when data are NMAR (*informative selection*)
- Need to specify a joint model for Y, X, and S
  - Lipsitz et al. (1999) suggest a EM approach, with missing values X integrated out given Y, S and other covariates. This EM method can handle X following any generalized linear model (including ordinal).
  - We handle violation of the MAR assumption by allowing the residuals for different models to be correlated through shared random effects, similar to Wu and Carroll (1988) for missing Y and the models for sample selection and endogenous covariates suggested by Heckman (1979)
  - To our knowledge, such a model has not been proposed before in the context of an ordinal missing covariate.

Missing covariate case

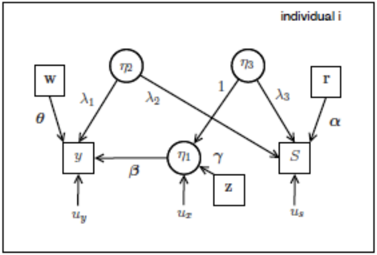


(b) Missing covariate

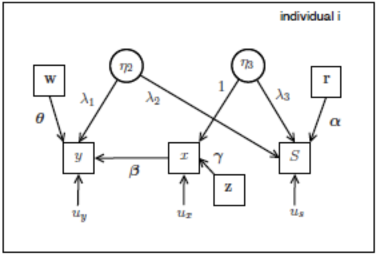
(c) Complete data

- Need to find discrete latent variable (unobserved) η_1 that can take the place of x when the ordinal explanatory variable is missing

Informative selection



(d) Missing covariate



(e) Complete data

Model for y

$$y_i = \begin{cases} \sum_{g=1}^{G} \beta_g 1(x_i = g) + \mathbf{w}_i'\boldsymbol{\theta} + \epsilon_{yi} & \text{if } x_i \text{ is observed} \\ \eta_{1i} + \mathbf{w}_i'\boldsymbol{\theta} + \epsilon_{yi} & \text{otherwise} \end{cases} \quad (1)$$

- $1(x_i = g)$ is a dummy variable for g-th of x_i with regression coefficient β_g
- η_{1i} is a discrete latent variable [Little and Schluchter, 1985] with

  $$P(\eta_{1i} = \beta_g | \mathbf{z}) = P(x_i = g | \mathbf{z}) \text{ in "latent class" } g$$

- $\mathbf{w}_i$ are other explanatory variables with coefficient $\boldsymbol{\theta}$

Model for missing covariate x

- Ordinal probit model with latent response x_i^*,

$$x_i^* \;=\; \mathbf{z}_i'\gamma + \epsilon_{xi}, \tag{2}$$

  - $x_i = g$ if $\kappa_{g-1} \leq x_i^* < \kappa_g$, $\{g = 1, \ldots, G\}$ and κ_g are threshold or cut-point parameters with $\kappa_0 = -\infty$ and $\kappa_G = \infty$.
  - $\mathbf{z}_i$ are explanatory variables with regression coefficients γ

- Latent variable η_{1i} is discrete with the conditional probabilities that $\eta_{1i} = \beta_g$ set equal to the conditional probabilities that $x_i = g$

Model for selection S

- Binary probit model with latent response S_i^*

$$S_i^* \;=\; \mathbf{r}_i'\boldsymbol{\alpha} + \epsilon_{si} \tag{3}$$

  - $S_i = 1(S_i^* > 0)$.
  - $\mathbf{r}_i$ are explanatory variables with regression coefficients $\boldsymbol{\alpha}$.

Errors, correlations

- Shared continuous latent variables η_{2i} and η_{3i} to make selection endogenous:

$$\begin{aligned} \epsilon_{yi} &= \lambda_1\eta_{2i} + u_{yi} \\ \epsilon_{xi} &= \eta_{3i} + u_{xi} \\ \epsilon_{Si} &= \lambda_2\eta_{2i} + \lambda_3\eta_{3i} + u_{Si}, \end{aligned} \tag{4}$$

[Heckman, 1979; Wu and Carroll, 1988]

  - η_{2i}, η_{3i}, u_{xi}, u_{Si} i.i.d. $N(0,1)$
  - $u_{yi} \sim N(0,\sigma^2)$
  - $\sigma^2 = 0.04$

$$\mathrm{Cor}(\epsilon_{yi},\epsilon_{Si}) = \frac{\lambda_1\lambda_2}{\sqrt{(\lambda_1^2+\sigma^2)(\lambda_2^2+\lambda_3^2+\sigma^2)}},$$

$$\mathrm{Cor}(\epsilon_{xi},\epsilon_{Si}) = \frac{\lambda_3}{\sqrt{(1+\sigma^2)(\lambda_2^2+\lambda_3^2+\sigma^2)}}.$$

Log-likelihood

$$\sum_{i,\ x_i\text{o},\ S_i=1} \ln\left\{\iint P_S(1|\eta_{2i},\eta_{3i})P_x(x_i|\eta_{3i})\phi_{x_i\text{o}}\left(y_i|x_i,\eta_{2i}\right)d\eta_{2i}d\eta_{3i}\right\}$$

$$+\sum_{i,\ x_i\bar{\text{o}},\ S_i=0} \ln\left\{\iint P_S(0|\eta_{2i},\eta_{3i})\left[\sum_{g=1}^{G} P_{\eta_1}(\beta_g|\eta_{3i})\phi_{x_i\bar{\text{o}}}\left(y_i|\beta_g,\eta_{2i}\right)\right]d\eta_{2i}d\eta_{3i}\right\}$$

$$+\sum_{i,\ x_i\bar{\text{o}},\ S_i\bar{\text{o}}} \ln\left\{\iint\left[\sum_{g=1}^{G} P_{\eta_1}(\beta_g|\eta_{3i})\phi_{x_i\bar{\text{o}}}\left(y_i|\beta_g,\eta_{2i}\right)\right]d\eta_{2i}d\eta_{3i}\right\}$$

Probabilities/densities

	y_i	x_i or η_{1i}	S_i
$x_i\text{o}$	$\phi_{x_i\text{o}}\left(y_i|x_i,\eta_{2i}\right)$	$P_x(x_i|\eta_{3i})$	$P_S(1|\eta_{2i},\eta_{3i})$
$x_i\bar{\text{o}}$	$\phi_{x_i\bar{\text{o}}}\left(y_i|\beta_g,\eta_{2i}\right)$	$P_{\eta_1}(\beta_g|\eta_{3i})$	$P_S(0|\eta_{2i},\eta_{3i})$

† For 493 responders with mother *"not a member of the household"*, add fourth term, identical to second term but with $P_S(1|\eta_{2i},\eta_{3i})$ instead of $P_S(0|\eta_{2i},\eta_{3i})$

Estimation

- Maximum Simulated Likelihood
  - Asymptotically equivalent to Maximum likelihood (replications R should grow faster than square root of the sample size $\sqrt{N}$) (Gourieroux and Monfort, 1993)
- Analytical first derivatives and OPG approx. of the Hessian
- Halton sequences cover the (0,1) interval better and require fewer draws to achieve high precision than random samples from uniform distribution
  - We use 800 Halton draws in all our regressions. Adding more draws did not change coefficients or standard errors
- Program written in Stata/Mata

Illustration: Ethnic gaps in school achievement at age 16 in England

Research question

How large are the differences in pupil attainment among ethnic groups at age 16 after allowing for differences in social background variables?

Mother's education is often a *"missing control"* either because no such information is available (administrative records) or because of item non-response (surveys). This missing covariate is likely to be a *"confounder"* in the relationship between achievement and ethnic group, leading to a problem of omitted variable bias.

Previous findings on ethnic gaps in GSCE results

- Wilson et al. (2005)
  - Data: NPD, 2002.
  - Ethnic minorities outperform the White British Majority
  - Only BC score less than WB
  - Chinese and Indian are the best achieving
- Strand (2008)
  - Data: LSYPE, 2006.
  - Similar findings

Data

- **National Pupil Database** (NPD)
  - *Long* but *narrow*: KS4 scores and ethnicity for whole population of pupils in maintained schools is available. Key covariates (e.g., mother's education) are missing [pupil id available]
- **Longitudinal Study of Young People in England** (LSYPE)
  - *Short* but *wide*: Information for a random sample of year 9 pupils in 2004. A rich set of controls (including mother's education) are available [pupil id available]
- **2001 UK Census**
  - Lower layer super output area characteristics (e.g., social class, qualifications, population density, deprivation, ethnicity)
- **NPD, LSYPE, and Census can be linked**
  - Data combination allows to add covariate information for a subset of pupils in the NPD
  - Problem: Covariate from LSYPE is missing for most pupils in NPD

Key variables

- [y_i]: Main outcome variable, capped new style style GCSE score [range from 0 to 540]. Available for everyone
- [$\mathbf{w}_i$]: Main explanatory variable of interest, ethnic group, and other covariates. Available for everyone
- [x_i]: Key covariate, mother's education. Only observed for:
  - Individuals sampled into LSYPE
  - Survey & item responders **in Wave 1**
- [$\mathbf{z}_i$]: Predictors of mother's education. Available for everyone
- [S_i]: Selection indicator
  - $S_i = 1$ if survey & item responder: $x_i\text{o}$
  - $S_i = 0$ if survey & item non-responder: $x_i\bar{\text{o}}$
  - $S_i = .$ if not included in survey: $x_i\bar{\text{o}}$, $S_i\bar{\text{o}}$
- [$\mathbf{r}_i$]: Predictors of unit & item response

Selection Variable S_i

Category	Symbol	Value	Freq.	%NPD	%LSYPE
Not LSYPE sampled	$x_i\bar{o}$, $S_i\bar{o}$	missing	545,130	96.69	0
LSYPE sampled, respondent	$x_i o$, $S_i = 1$	1	13,372†	2.37	71.59
LSYPE sampled, non-respondent	$x_i\bar{o}$, $S_i = 0$	0	5,307	0.94	28.41
Total			563,809	100	100

† For 493 of these cases, x_i is missing although $S_i = 1$ because mother was reported to be *"not a member of the household"* but survey was otherwise completed.

Mothers' education, ordinal x_i

Category	Freq.	%	$\%^w$	$\bar{y}$	$\bar{y}^w$
1. No qualification	3,451	26.80	19.83	271.28	252.61
2. Other qualifications	1,215	9.43	10.5	278.60	272.24
3. GCSE grades A-C or equiv	3,869	30.04	33.49	302.82	298.69
4. GCE A level or equiv	1,586	12.31	13.63	323.21	321.88
5. Higher education no degree	1,539	11.95	12.54	333.54	333.22
6. Degree or equivalent	1,219	9.47	10.02	366.76	368.10
Total	12,879				

^{w}Statistic calculated using probability weights for the LSYPE.

Ethnic group

▸ Results 1

Category	Freq.	%	$\bar{y}$	S_i		x_i	
				$10\% S_i\text{o}$	$\% \frac{S_i=1}{S_i\text{o}}$	$\%(\geq 3)$	$\%(\geq 3)^w$
White british	461,070	81.78	298.47	20.46	73.65	73.48	73.35
White other	13,168	2.34	306.93	0.53	67.45	53.61	54.89
Mixed	12,596	2.23	294.99	1.91	70.34	67.99	69.31
Indian	13,061	2.32	334.88	2.10	72.76	46.67	47.95
Pakistani	13,083	2.32	288.33	2.14	68.69	20.67	21.14
Bangladeshi	5,516	0.98	297.92	1.65	68.14	10.54	10.77
Other asian	3,909	0.69	317.65	0.20	71.30	50.62	50.55
Caribbean	8,062	1.43	271.64	1.49	62.98	79.76	81.22
African	9,703	1.72	285.22	1.50	63.83	53.36	52.39
Other black	2,481	0.44	272.69	0.13	62.16	70.73	74.04
Chinese	2,028	0.36	361.65	0.09	50.94	32.00	31.59
Any other	4,931	0.87	285.57	0.23	67.44	32.53	28.87
Refused	6,545	1.16	297.44	0.27	68.39	82.18	83.25
No data	7,656	1.36	277.90	0.43	74.79	67.26	67.39
Total	563,809						

^{w}Statistic calculated using probability weights for the LSYPE.

Survey/item response is likely to be informative

- Selection mechanism S_i is endogenous with respect to both achievement y_i and mother's education x_i and therefore non ignorable [Rubin, 1967; Heckman, 1979] [Lipsitz et al., 1999]
    - Example 1: Mothers of high performers are more likely to be interested in child's education and co-operate with the school and the survey

      $\Rightarrow$ Positive correlation between y_i and S_i?
    - Example 2: Highly educated mothers are more likely to have tight schedules and therefore less willing/available to participate in the survey

      $\Rightarrow$ Negative correlation between x_i and S_i?
- After controlling for LSYPE design variables missingness of S_i is ignorable

Exclusion restrictions

- **LSYPE interviewer company is predictor of S_i but not of y_i or x_i**
  - British Market Research Bureau (lead contractor)
  - Ipsos MORI
  - GfK NOP
  - Joint work BMRB-Mori or NOP-Mori
- **Winter born dummy is predictor of y_i but not of x_i or S_i**
  - Due to Local Authority policy, a child born in the summer may enter school almost a year ealier than the eldest pupil in her/his cohort (Crawford et al. 2007, p. 2)

Variables in all equations

Table: Variables in all equations

Variable	Description	Reason
FSM dummy	Taking free school meal (No)	SES proxy from NPD
Deprived school dummy	Top quintile of %FSM (No)	Design variable
Ethnicity dummies	8 ethnicities (White)	Variable of main interest; design variable
School-type by gender dummies	4 groups: mixed/boys, mixed/girl, boys/boy, (girls/girl)	Predictor of selection
Geographic region dummies	9 regions (East Midlands)	Predictor of selection

Note. Category in brackets is the reference group.

Results for exclusion restrictions and selection

- **Model for** y_i

Variable	Est	(SE)
winterbn	.06	(.002)

- **Model for** S_i (BMRB is reference group)

Company	Est	(SE)
NOP	-.08	(.021)
MORI	-.17	(.035)
BMRB-Mori or NOP-Mori	-.71	(.112)

- **Correlations**

$$\widehat{\mathrm{Cor}}(\epsilon_{yi}, \epsilon_{Si}) = .16 \quad (.010)$$
$$\widehat{\mathrm{Cor}}(\epsilon_{xi}, \epsilon_{Si}) = -.23 \quad (.014)$$

Results for standardised capped GCSE point score

▸ descriptive statats

	linear regressions				Missing covariate model[a,d]			
	NPD[b]		LSYPE[a,c]		Benchmark		Extra controls	
Variable	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Ethnic group (White British)								
White other	0.116‡	0.008	0.384‡	0.059	0.159‡	0.006	0.129‡	0.006
Mixed	0.054‡	0.009	0.023	0.040	0.072‡	0.005	0.072‡	0.005
Indian	0.415‡	0.009	0.513‡	0.033	0.388‡	0.005	0.345‡	0.005
Pakistani	0.232‡	0.009	0.468‡	0.041	0.358‡	0.005	0.342‡	0.005
Bangladeshi	0.407‡	0.013	0.686‡	0.051	1.608‡	0.006	0.697‡	0.007
Asian other	0.282‡	0.015	0.326†	0.110	0.270‡	0.010	0.250‡	0.010
Black C.	-0.106‡	0.011	-0.201‡	0.049	-0.183‡	0.007	-0.170‡	0.007
Black A.	0.122‡	0.010	0.232‡	0.053	0.114‡	0.006	0.111‡	0.006
Black other	-0.086‡	0.019	-0.187†	0.146	-0.144‡	0.015	-0.125‡	0.014
Chinese	0.606‡	0.021	0.860‡	0.159	0.473‡	0.014	0.467‡	0.013
Any other	0.095‡	0.014	0.464‡	0.104	0.167‡	0.009	0.141‡	0.009
Refused	-0.028†	0.012	-0.113	0.105	-0.009‡	0.009	-0.022‡	0.009
No data	-0.223‡	0.011	-0.056	0.075	-0.100‡	0.008	-0.094‡	0.008

Results for standardised capped GCSE point score

▸ descriptive statats

	linear regressions				Missing covariate model[a,d]			
	NPD[b]		LSYPE[a,c]		Benchmark		Extra controls	
Variable	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Mother education								
No qual.			-0.142	0.054	-1.415‡	0.005	-1.369‡	0.018
Other qual.			0.039	0.054	0.392‡	0.008	0.411‡	0.019
GCSE A-C			0.281‡	0.049	0.517‡	0.006	0.509‡	0.018
GCE A level			0.466‡	0.050	0.570‡	0.008	0.553‡	0.019
Some HE			0.565‡	0.053	0.589‡	0.007	0.572‡	0.019
Degree			0.870‡	0.051	0.646‡	0.008	0.648‡	0.019

Note: ‡(†) Significant at 1% (5%). OPG standard errors reported. Dependent variable is the standardised capped new style GCSE score. (a) To ease comparison across columns these models have no constant term to ensure that coefficients on mother's education can be interpreted as the mean when other controls are zero. The coefficients on mother's education are also the locations of the discrete latent variable η_{1i}. (b) Ordinary least squares regression (c) Weighted least squares regression. (d) Details on coefficients in selection and missing covariate equations are given in Table 9 of the paper.

Discussion

- Ethnic gap estimates increase after controlling for mother's education

  ⇒ Cannot ignore mother's education

- Selection is informative

  ⇒ Cannot use listwise deletion, with LSYPE data only
  ⇒ Cannot use multiple imputation, with merged data

- Standard errors smaller for merged data than for LSYPE

  ⇒ Should not apply model only to pupils sampled into LSYPE (excluding $S_i\bar{o}$)

References

- Crawford, C., Dearden, L., Meghir, C., 2007. When you are born matters: the impact of date of birth on educational outcomes in England. Center for the Economics of Education dicussion paper No. 93.
- Heckman, J. J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.
- Lipsitz, S. R., Ibrahim, J. G., Chen, M.-H., H. Peterson, H., 1999. Non-ignorable missing covariates in generalized linear models. Statistics in Medicine 18, 2435–2448.
- Miranda, A., Rabe-Hesketh, S., 2010. Missing ordinal covariates with informative selection. DoQSS working papers No. 10-16.
- Little, R. J. A., Schluchter, M., 1985. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72, 497–512.
- Rubin, D. B., 1976. Inference and missing data. Biometrika 63, 581–592.
- Strand, S., 2008. Minority ethnic pupils in the Longitudinal Study of Young People in England: Extension report on performance in public examinations at age 16, Tech. rep. DCSF-RR029
- Wilson, D., Burgess, S., Briggs, A., 2005. The dynamics of school attainment of England's ethnic minorities, Tech. rep. CMPO 05/130, University of Bristol.