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

- \([y]\): main response variable (continuous)
- \([x]\): key ordinal control variable
- \([w_i]\): predictors for \(y\)
- \([z_i]\): predictors of \(x\)

Note: \(w_i = \{z_i, y\text{-only explanatory variables}\}\)
Missing covariate case

- $x$ 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,

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


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

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 \( \hat{Pr}(X|Y, V) \). 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) — $Pr(S \mid Y, X, V)$

- 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

- Need to find discrete latent variable (unobserved) $\eta_1$ that can take the place of $x$ when the ordinal explanatory variable is missing
Informative selection
Model for $y$

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

- $1(x_i = g)$ is a dummy variable for $g$-th of $x_i$ with regression coefficient $\beta_g$
- $\eta_{1i}$ is a discrete latent variable [Little and Schluchter, 1985] with $P(\eta_{1i} = \beta_g | z) = P(x_i = g | z)$ in “latent class” $g$
- $w_i$ are other explanatory variables with coefficient $\theta$
Model for missing covariate $x$

- Ordinal probit model with latent response $x^*_i$,

$$x^*_i = z'_i \gamma + \epsilon_{xi},$$

(2)

- $x_i = g$ if $\kappa_{g-1} \leq x^*_i < \kappa_g$, $\{g = 1, \ldots, G\}$ and $\kappa_g$ are threshold or cut-point parameters with $\kappa_0 = -\infty$ and $\kappa_G = \infty$.

- $z_i$ are explanatory variables with regression coefficients $\gamma$

- Latent variable $\eta_{1i}$ 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 $S_i^*$

$$S_i^* = r_i'\alpha + \epsilon_{si} \quad (3)$$

- $S_i = 1(S_i^* > 0)$.
- $r_i$ are explanatory variables with regression coefficients $\alpha$. 
Errors, correlations

- Shared continuous latent variables $\eta_2i$ and $\eta_3i$ to make selection endogenous:

\[
\begin{align*}
\epsilon_{yi} & = \lambda_1 \eta_{2i} + u_{yi} \\
\epsilon_{xi} & = \eta_{3i} + u_{xi} \\
\epsilon_{Si} & = \lambda_2 \eta_{2i} + \lambda_3 \eta_{3i} + u_{Si},
\end{align*}
\]

(4)

[Heckman, 1979; Wu and Carroll, 1988]

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

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

\[
\text{Cor}(\epsilon_{xi}, \epsilon_{Si}) = \frac{\lambda_3}{\sqrt{(1 + \sigma^2)(\lambda_2^2 + \lambda_3^2 + \sigma^2)}}.
\]
Log-likelihood

\[
\sum_{i, x_{i\theta}, S_i=1} \ln \left\{ \int \int P_S(1|\eta_2i, \eta_3i) P_x(x_i|\eta_3i) \phi_{x_{i\theta}}(y_i|x_i, \eta_2i) \, d\eta_2i \, d\eta_3i \right\} \\
+ \sum_{i, x_{i\bar{\theta}}, S_i=0} \ln \left\{ \int \int P_S(0|\eta_2i, \eta_3i) \left[ \sum_{g=1}^{G} P_{\eta_1}(\beta_g|\eta_3i) \phi_{x_{i\bar{\theta}}}(y_i|\beta_g, \eta_2i) \right] \, d\eta_2i \, d\eta_3i \right\} \\
+ \sum_{i, x_{i\bar{\theta}}, S_i\bar{\theta}} \ln \left\{ \int \int \left[ \sum_{g=1}^{G} P_{\eta_1}(\beta_g|\eta_3i) \phi_{x_{i\bar{\theta}}}(y_i|\beta_g, \eta_2i) \right] \, d\eta_2i \, d\eta_3i \right\}
\]

Probabilities/densities

<table>
<thead>
<tr>
<th></th>
<th>(y_i)</th>
<th>(x_i) or (\eta_{1i})</th>
<th>(S_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x_{i\theta})</td>
<td>(\phi_{x_{i\theta}}(y_i</td>
<td>x_i, \eta_{2i}))</td>
<td>(P_x(x_i</td>
</tr>
<tr>
<td>(x_{i\bar{\theta}})</td>
<td>(\phi_{x_{i\bar{\theta}}}(y_i</td>
<td>\beta_g, \eta_{2i}))</td>
<td>(P_{\eta_1}(\beta_g</td>
</tr>
</tbody>
</table>

* 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)
  - 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 GCSE score [range from 0 to 540]. Available for everyone

- \([w_i]\): 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

- \([z_i]\): Predictors of mother’s education. Available for everyone

- \([S_i]\): Selection indicator
  - \(S_i = 1\) if survey & item responder: \(x_i\circ\)
  - \(S_i = 0\) if survey & item non-responder: \(x_i\bar{\circ}\)
  - \(S_i = .\) if not included in survey: \(x_i\bar{\circ}, S_i\bar{\circ}\)

- \([r_i]\): Predictors of unit & item response
## Selection Variable $S_i$

<table>
<thead>
<tr>
<th>Category</th>
<th>Symbol</th>
<th>Value</th>
<th>Freq.</th>
<th>%NPD</th>
<th>%LSYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not LSYPE sampled</td>
<td>$x_i \bar{\sigma}, S_i \bar{\sigma}$</td>
<td>missing</td>
<td>545,130</td>
<td>96.69</td>
<td>0</td>
</tr>
<tr>
<td>LSYPE sampled, respondent</td>
<td>$x_i \sigma, S_i = 1$</td>
<td>1</td>
<td>13,372†</td>
<td>2.37</td>
<td>71.59</td>
</tr>
<tr>
<td>LSYPE sampled, non-respondent</td>
<td>$x_i \bar{\sigma}, S_i = 0$</td>
<td>0</td>
<td>5,307</td>
<td>0.94</td>
<td>28.41</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>563,809</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

† 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.
Descriptive statistics

Mothers’ education, ordinal $x_i$

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

$^w$Statistic calculated using probability weights for the LSYPE.
# Ethnic group

<table>
<thead>
<tr>
<th>Category</th>
<th>Freq.</th>
<th>%</th>
<th>$\bar{y}$</th>
<th>$S_i$</th>
<th>$%\left(S_{i0}\right)$</th>
<th>$%\left(S_{i0}^{-1}\right)$</th>
<th>$%\left(\geq 3\right)$</th>
<th>$%\left(\geq 3\right)^{w}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>White british</td>
<td>461,070</td>
<td>81.78</td>
<td>298.47</td>
<td>20.46</td>
<td>73.65</td>
<td>73.48</td>
<td>73.35</td>
<td></td>
</tr>
<tr>
<td>White other</td>
<td>13,168</td>
<td>2.34</td>
<td>306.93</td>
<td>0.53</td>
<td>67.45</td>
<td>53.61</td>
<td>54.89</td>
<td></td>
</tr>
<tr>
<td>Mixed</td>
<td>12,596</td>
<td>2.23</td>
<td>294.99</td>
<td>1.91</td>
<td>70.34</td>
<td>67.99</td>
<td>69.31</td>
<td></td>
</tr>
<tr>
<td>Indian</td>
<td>13,061</td>
<td>2.32</td>
<td>334.88</td>
<td>2.10</td>
<td>72.76</td>
<td>46.67</td>
<td>47.95</td>
<td></td>
</tr>
<tr>
<td>Pakistani</td>
<td>13,083</td>
<td>2.32</td>
<td>288.33</td>
<td>2.14</td>
<td>68.69</td>
<td>20.67</td>
<td>21.14</td>
<td></td>
</tr>
<tr>
<td>Bangladeshi</td>
<td>5,516</td>
<td>0.98</td>
<td>297.92</td>
<td>1.65</td>
<td>68.14</td>
<td>10.54</td>
<td>10.77</td>
<td></td>
</tr>
<tr>
<td>Other asian</td>
<td>3,909</td>
<td>0.69</td>
<td>317.65</td>
<td>0.20</td>
<td>71.30</td>
<td>50.62</td>
<td>50.55</td>
<td></td>
</tr>
<tr>
<td>Caribbean</td>
<td>8,062</td>
<td>1.43</td>
<td>271.64</td>
<td>1.49</td>
<td>62.98</td>
<td>79.76</td>
<td>81.22</td>
<td></td>
</tr>
<tr>
<td>African</td>
<td>9,703</td>
<td>1.72</td>
<td>285.22</td>
<td>1.50</td>
<td>63.83</td>
<td>53.36</td>
<td>52.39</td>
<td></td>
</tr>
<tr>
<td>Other black</td>
<td>2,481</td>
<td>0.44</td>
<td>272.69</td>
<td>0.13</td>
<td>62.16</td>
<td>70.73</td>
<td>74.04</td>
<td></td>
</tr>
<tr>
<td>Chinese</td>
<td>2,028</td>
<td>0.36</td>
<td>361.65</td>
<td>0.09</td>
<td>50.94</td>
<td>32.00</td>
<td>31.59</td>
<td></td>
</tr>
<tr>
<td>Any other</td>
<td>4,931</td>
<td>0.87</td>
<td>285.57</td>
<td>0.23</td>
<td>67.44</td>
<td>32.53</td>
<td>28.87</td>
<td></td>
</tr>
<tr>
<td>Refused</td>
<td>6,545</td>
<td>1.16</td>
<td>297.44</td>
<td>0.27</td>
<td>68.39</td>
<td>82.18</td>
<td>83.25</td>
<td></td>
</tr>
<tr>
<td>No data</td>
<td>7,656</td>
<td>1.36</td>
<td>277.90</td>
<td>0.43</td>
<td>74.79</td>
<td>67.26</td>
<td>67.39</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>563,809</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^{w}$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
  ⇒ 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
  ⇒ 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 earlier than the eldest pupil in her/his cohort (Crawford et al. 2007, p. 2)
### Variables in all equations

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

Note. Category in brackets is the reference group.
Results for exclusion restrictions and selection

- **Model for** $y_i$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est</th>
<th>(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>winterbn</td>
<td>.06</td>
<td>(.002)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Company</th>
<th>Est</th>
<th>(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOP</td>
<td>-.08</td>
<td>(.021)</td>
</tr>
<tr>
<td>MORI</td>
<td>-.17</td>
<td>(.035)</td>
</tr>
<tr>
<td>BMRB-Mori or NOP-Mori</td>
<td>-.71</td>
<td>(.112)</td>
</tr>
</tbody>
</table>

- **Correlations**

  \[
  \hat{\text{Cor}}(\epsilon_{yi}, \epsilon_{Si}) = .16 \ (0.010) \\
  \hat{\text{Cor}}(\epsilon_{xi}, \epsilon_{Si}) = -.23 \ (0.014) 
  \]
## Results for standardised capped GCSE point score

<table>
<thead>
<tr>
<th>Ethnic group (White British)</th>
<th>Linear regressions</th>
<th>Missing covariate model&lt;sup&gt;a,d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NPD&lt;sup&gt;b&lt;/sup&gt;</td>
<td>LSYPE&lt;sup&gt;a,c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Coeff.  SE</td>
<td>Coeff.  SE</td>
</tr>
<tr>
<td>White other</td>
<td>0.116&lt;sup&gt;†&lt;/sup&gt; 0.008</td>
<td>0.384&lt;sup&gt;†&lt;/sup&gt; 0.059</td>
</tr>
<tr>
<td>Mixed</td>
<td>0.054&lt;sup&gt;†&lt;/sup&gt; 0.009</td>
<td>0.023 0.040</td>
</tr>
<tr>
<td>Indian</td>
<td>0.415&lt;sup&gt;†&lt;/sup&gt; 0.009</td>
<td>0.513&lt;sup&gt;†&lt;/sup&gt; 0.033</td>
</tr>
<tr>
<td>Pakistani</td>
<td>0.232&lt;sup&gt;†&lt;/sup&gt; 0.009</td>
<td>0.468&lt;sup&gt;†&lt;/sup&gt; 0.041</td>
</tr>
<tr>
<td>Bangladeshi</td>
<td>0.407&lt;sup&gt;†&lt;/sup&gt; 0.013</td>
<td>0.686&lt;sup&gt;†&lt;/sup&gt; 0.051</td>
</tr>
<tr>
<td>Asian other</td>
<td>0.282&lt;sup&gt;†&lt;/sup&gt; 0.015</td>
<td>0.326&lt;sup&gt;†&lt;/sup&gt; 0.110</td>
</tr>
<tr>
<td>Black C.</td>
<td>-0.106&lt;sup&gt;†&lt;/sup&gt; 0.011</td>
<td>-0.201&lt;sup&gt;†&lt;/sup&gt; 0.049</td>
</tr>
<tr>
<td>Black A.</td>
<td>0.122&lt;sup&gt;†&lt;/sup&gt; 0.010</td>
<td>0.232&lt;sup&gt;†&lt;/sup&gt; 0.053</td>
</tr>
<tr>
<td>Black other</td>
<td>-0.086&lt;sup&gt;†&lt;/sup&gt; 0.019</td>
<td>-0.187&lt;sup&gt;†&lt;/sup&gt; 0.146</td>
</tr>
<tr>
<td>Chinese</td>
<td>0.606&lt;sup&gt;†&lt;/sup&gt; 0.021</td>
<td>0.860&lt;sup&gt;†&lt;/sup&gt; 0.159</td>
</tr>
<tr>
<td>Any other</td>
<td>0.095&lt;sup&gt;†&lt;/sup&gt; 0.014</td>
<td>0.464&lt;sup&gt;†&lt;/sup&gt; 0.104</td>
</tr>
<tr>
<td>Refused</td>
<td>-0.028&lt;sup&gt;†&lt;/sup&gt; 0.012</td>
<td>-0.113 0.105</td>
</tr>
<tr>
<td>No data</td>
<td>-0.223&lt;sup&gt;†&lt;/sup&gt; 0.011</td>
<td>-0.056 0.075</td>
</tr>
</tbody>
</table>
## Results for standardised capped GCSE point score

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
<th>Coeff.</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mother education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No qual.</td>
<td>-0.142</td>
<td>0.054</td>
<td>-1.415</td>
<td>‡</td>
<td>0.005</td>
<td>‡</td>
<td>-1.369</td>
<td>‡</td>
</tr>
<tr>
<td>Other qual.</td>
<td>0.039</td>
<td>0.054</td>
<td>0.392</td>
<td>‡</td>
<td>0.008</td>
<td>‡</td>
<td>0.411</td>
<td>‡</td>
</tr>
<tr>
<td>GCSE A-C</td>
<td>0.281†</td>
<td>0.049</td>
<td>0.517†</td>
<td>0.006</td>
<td>0.509†</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCE A level</td>
<td>0.466†</td>
<td>0.050</td>
<td>0.570†</td>
<td>0.008</td>
<td>0.553†</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some HE</td>
<td>0.565†</td>
<td>0.053</td>
<td>0.589†</td>
<td>0.007</td>
<td>0.572†</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Degree</td>
<td>0.870†</td>
<td>0.051</td>
<td>0.646†</td>
<td>0.008</td>
<td>0.648†</td>
<td>0.019</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**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 $\eta_{1i}$. (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
  \[\Rightarrow\text{Cannot ignore mother’s education}\]

- Selection is informative
  \[\Rightarrow\text{Cannot use listwise deletion, with LSYPE data only}\]
  \[\Rightarrow\text{Cannot use multiple imputation, with merged data}\]

- Standard errors smaller for merged data than for LSYPE
  \[\Rightarrow\text{Should not apply model only to pupils sampled into LSYPE (excluding } S_i \bar{o})\]
References


