# Missing Covariates with Informative Selection

Alfonso Miranda & Sophia Rabe-Hesketh

ADMIN Node

# 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): NPD/PLASC/Census, 2005. (TPS). [Chinese, Indian, Other, White, Bangladeshi, Black African, Pakistani, Black other, Black Caribbean] Large changes when controlling for covariates
- Connolly (2006): Youth Cohort Study of England and Wales, 1999. (TPS & 5A*-C). [Chinese, Indian, White, Black, Pakistani, Bangladeshi]
- Rothon (2007): Youth Cohort Study of England and Wales, 1991-2000. (5A*-C). Controlling for social class: [Indian, White, Black, Pakistani]
- Patacchini and Zenou (2009): National Child Development Study, 1974. (Maths/Reading scores) Relationship between parental involvement and Black African - White gap
- Strand (2008): Longitudinal Study of Young People in England, 2006. (5A*-C & TPS) [Indian, Other, White, Mixed, Bangladeshi, Black African, Pakistani, Black Caribbean]

TPS: Total Point Score for GCSEs (continuous)

5A*-C: At least 5 GCSEs with grades A* to C (binary)

# Data



- National Pupil Data Base (NPD) is an administrative data base containing records for the whole population of pupils (excluding private schools) in England from 2002 onwards, covering both pupil's characteristics and their examination results. Limited information on family income and socioeconomic status.
  - Census data and other datasets with area-level variables can be merged with NPD.
- Longitudinal Survey of Young People in England (LSYPE) is a longitudinal survey of a random sample of Year 9 pupils in 2004 and their parents in England. Interviews are conducted annually. Survey contains detailed information on family income, socioeconomic status, parents' education.
  - LSYPE can be merged with NPD.

# Exploiting data linkage



- Two sources of information
  - NPD is *long* but *narrow*. Information for whole population is available but key covariates (e.g., mother's education) are missing.
  - LSYPE is *short* but *wide*. Information only for a random sample but a rich set of controls are available.
- Link NPD and LSYPE: to add covariate information for a subset of pupils in the NPD.
- Problem: Covariate from LSYPE missing for most pupils in NPD

# Samples/design



- LSYPE
  - Pupils in in Year 9 in England in 2004 (Wave 1), age 16 in 2006 (Wave 3)

    We exclude pupils from non-maintained schools
  - Two-stage design
    1. Schools: Oversample top quintile in %FSM, "taking into account number of pupils from different minority groups"
    2. Pupils: Oversample major ethnic minority groups to achieve target issued samples of 1000 per group
- NPD and merged data
  - Pupils who took GCSEs (Key stage 4) in 2006,

    We exclude pupils
    - from non-maintained schools
    - from Wales
    - with Special Educational Needs (SEN): statemented
    - with missing GCSE score

# Key variables



- $[y_i]$: Main outcome variable, GCSE score 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, is observed only 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: $x_i$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 survey & item response

# Descriptive statistics



Table: 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$^\dagger$ | 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 |

$^\dagger$ 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



Table: Capped GCSE new style point score $y_i$

| Category | Symbol | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Not LSYPE sampled | $x_i\bar{o}$, $S_i\bar{o}$ | 298.40 | 101.87 | 0 | 540 |
| LSYPE sampled, respondent | $x_i o$, $S_i = 1$ | 302.46 | 98.54 | 0 | 502 |
| LSYPE sampled, no respondent | $x_i\bar{o}$, $S_i = 0$ | 290.10 | 103.35 | 0 | 483 |

# Descriptive statistics



Table: Mothers' education, ordinal $x_i$

| Category | Freq. | % | Cum. | $\bar{y}$ |
|---|---|---|---|---|
| 1. No qualification | 3,451 | 26.80 | 276.80 | 271.28 |
| 2. Other qualifications | 1,215 | 9.43 | 36.23 | 278.60 |
| 3. GCSE grades A-C or equiv | 3,869 | 30.04 | 66.27 | 302.82 |
| 4. GCE A level or equiv | 1,586 | 12.31 | 78.59 | 323.21 |
| 5. Higher education no degree | 1,539 | 11.95 | 90.53 | 333.54 |
| 6. Degree or equivalent | 1,219 | 9.47 | 100 | 366.76 |
| Total | 12,877 | 100 | | |

Is the ordering for 1. and 2. correct/important?

# Descriptive statistics



Table: Ethnic group

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

# Informative Selection



- Survey/item response is likely to be endogenous or informative or non-ignorable: Related to both achievement $y_i$ [Rubin, 1967; Heckman, 1979] and mother's education $x_i$ [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 (that determined sampling probabilities), missingness of $S_i$ is ignorable

# Model for achievement: y



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

- $1(x_i = a_g)$ is a dummy variable for $g$th value $a_g$ of $x_i$ with regression coefficient $\beta_g$
- $\mathbf{w}_i$ are other explanatory variables, including ethnic group, with regression coefficients $\boldsymbol{\beta}_{G+1}$
- $\eta_{1i}$ is a discrete latent variable [Little and Schluchter, 1985]

$$\eta_{1i} = \beta_g \text{ in "latent class" } g$$

# Model for achievement: y



## Exclusion restriction 1

- **Summer vs. winter born** enters model for $y_i$ and not models for $x_i$ and $S_i$.

  According to English law, children must have started school by the beginning of the term (January, April, or September) following their fifth birthday but no minimum age is specified.

  - Children born in the summer enter school in the January or April, 1 to 2 terms before their fifth birthday
  - Children born in the autumn start in September, close to their fifth birthday
    (see, for instance, Dearden, Crawford, Meghir 2007).

# Model for mother's education: x



- Ordinal probit model with latent response $x_i^*$,

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

  - $x_i = a_g$ 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 $\kappa_G = \infty$.
  - $\mathbf{z}_i$ 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 $x_i = a_g$.

# Model for survey/item response: 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 $\boldsymbol{\alpha}$.

## Exclusion restriction 2

**Company that did LSYPE field work in Wave 1** enters model for $S_i$ but not models for $x_i$ and $y_i$.

3 companies and 4 goups: (a) British Market Research Bureau; (b) Ipsos MORI; (c) GfK NOP; (d) joint work BMRB-Mori or NOP-Mori. Companies may differ in their ability, effort, or incentives to track down and interview individuals

Table: Company doing LSYPE field work

| Category | Freq. | % | %S=1 |
|---|---|---|---|
| BMRB | 8,061 | 43.16 | 73.63 |
| NOP | 8,316 | 44.52 | 71.90 |
| Mori | 2,183 | 11.69 | 64.64 |
| BMRB-Mori or NOP-Mori | 119 | 0.64 | 39.50 |
| Total | 18,679 | 100 | |

# 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.

# Model: errors, correlations



- Shared latent variables $\eta_{2i}$ and $\eta_{3i}$ to make selection endogenous:

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

[Heckman, 1979; Wu and Carroll, 1988]

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

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

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

# Log-likelihood



- Log-likelihood$^{\dagger}$:

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

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

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

Probabilities/densities

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

$^{\dagger}$ 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
- 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
- Program written in Stata/Mata
- Really fast!
  - Stata 10/MP + 12 processors + 100 Halton draws + 563,658 obs = 7hrs
  - Stata 10/MP + 12 processors + 800 Halton draws + 563,658 obs = 25hrs

# Results for exclusion restrictions and selection



- Model for $y_i$, exclusion restriction 1

| Variable | Est | (SE) |
|---|---|---|
| winterbn | .06 | (.003) |

- Model for $S_i$, exclusion restriction 2 (BMRB is reference group)

| Company | Est | (SE) |
|---|---|---|
| NOP | -.06 | (.021) |
| MORI | -.17 | (.033) |
| BMRB-Mori or NOP-Mori | -.72 | (.111) |

- Correlations, both highly significant:

$$\widehat{\mathrm{Cor}}(\epsilon_{yi}, \epsilon_{Si}) = 0.16$$
$$\widehat{\mathrm{Cor}}(\epsilon_{xi}, \epsilon_{Si}) = -0.22$$

# Results for standardised capped GCSE new style point score



| | NPD | | Merged | | LSYPE | |
|---|---|---|---|---|---|---|
| Category | Est | (SE) | Est | (SE) | Est | (SE) |
| Mixed (white) | .02 | (.009) | .02 | (.009) | .01 | (.035) |
| Indian | .34 | (.009) | .41 | (.009) | .43 | (.033) |
| Pakistani | .09 | (.009) | .21 | (.010) | .31 | (.035) |
| Bangladeshi | .27 | (.013) | .34 | (.014) | .54 | (.042) |
| Caribbean | -.22 | (.011) | -.14 | (.010) | -.28 | (.043) |
| African | .00 | (.010) | .09 | (.010) | .11 | (.044) |
| Other | .13 | (.009) | .23 | (.010) | .22 | (.060) |
| Refused | -.01 | (.012) | -.02 | (.017) | -.09 | (.089) |
| No data | -.23 | (.011) | -.16 | (.015) | -.11 | (.070) |
| No qual. | | | .38 | (.010) | -.39 | (.033) |
| Other qual. | | | -1.43 | (.010) | -.24 | (.036) |
| GCSE A-C | | | .46 | (.010) | .00 | (.030) |
| GCE A level | | | .48 | (.013) | .18 | (.035) |
| Some higher ed. | | | .51 | (.010) | .28 | (.035) |
| Degree | | | .57 | (.013) | .58 | (.037) |

# Discussion



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

  $\Rightarrow$ Cannot ignore mother's education

- Selection is informative

  $\Rightarrow$ Cannot use listwise deletion, with LSYPE data only
  $\Rightarrow$ Cannot use multiple imputation, with merged data

- Standard errors smaller for merged data than for LSYPE

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

# Next steps



- Same model, but only for pupils sampled into LSYPE (exclude $S_i\bar{o}$)
- Include super output area census variables
  - Better control for background in model for $y$
  - Better predict mother's education $x$
  - Better predict sample selection $S$
    - Population density, IDACI, qualifications, country of birth, unemployment, income support, ?
- Candidates for extra school variables from PLASC
  - % FSM, pupil/teacher ratio, ?
- Cluster standard errors at school level?

# References




- Connolly, P., 2006. Summary statistics, educational achievement gaps and the ecological fallacy. Oxford Review of Education 32, 235–252.
- 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.
- Little, R. J. A., Schluchter, M., 1985. Maximum likelihood estimation for mixed continuous and categorical data with missing values. Biometrika 72, 497–512.
- Patacchini, E., Zenou, Y., 2009. On the sources of the black-white test score gap in europe. Economics Letters 102, 49–52.
- Rothon, C., 2007. Can achievement differentials be explained by social class alone? Ethnicities 7, 306–322.
- 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.
- Wu, M. C., Carroll, R. J., 1988. Estimation and comparison of change in the presence of informative right censoring by modeling the censoring process. Biometrics 44, 175–188.