Analyzing Inequality of Educational Opportunities using Stacked Surveys with Missing Data

Maarten L. Buis

Vrije Universiteit Amsterdam

Department of Social Research Methodology

http://home.fsw.vu.nl/m.buis

Outline

- baseline model
- Missing Data
  - Multiple Imputation of multiple surveys
  - assess plausibility of results
- Nesting within surveys
  - Random effects model
  - assess plausibility of results

Potential problems:

- Missing data (11,000 out of 99,000 cases are missing)
  - Potential bias
  - Not as efficient as could be

Potential problems:

- Missing data (11,000 out of 99,000 cases are missing)
  - Potential bias
  - Not as efficient as could be
- Individuals are nested in surveys (50 surveys)
  - Potential bias
  - Too efficient

Conclusions

- Missing data
  - Virtually no bias was found.
  - Virtually no gain in power was achieved by using Multiple Imputation.
- Nested structure of the data
  - Outlying studies have lead to an underestimation of the trend in IEO in pooled regression.
  - Standard errors increases a little when controlling for nested structure.

Baseline model

- Linear regression of highest achieved level of education on:

Baseline model

- Linear regression of highest achieved level of education on:
  - father's occupational status ($status$), which captures the Inequality of Educational Opportunity (IEO),

Baseline model

- Linear regression of highest achieved level of education on:
  - father's occupational status ($status$), which captures the Inequality of Educational Opportunity (IEO),
  - year of child's birth ($birthyear$), which captures educational expansion, and is added as a spline with three equally spaced knots to allow for non-linearity,

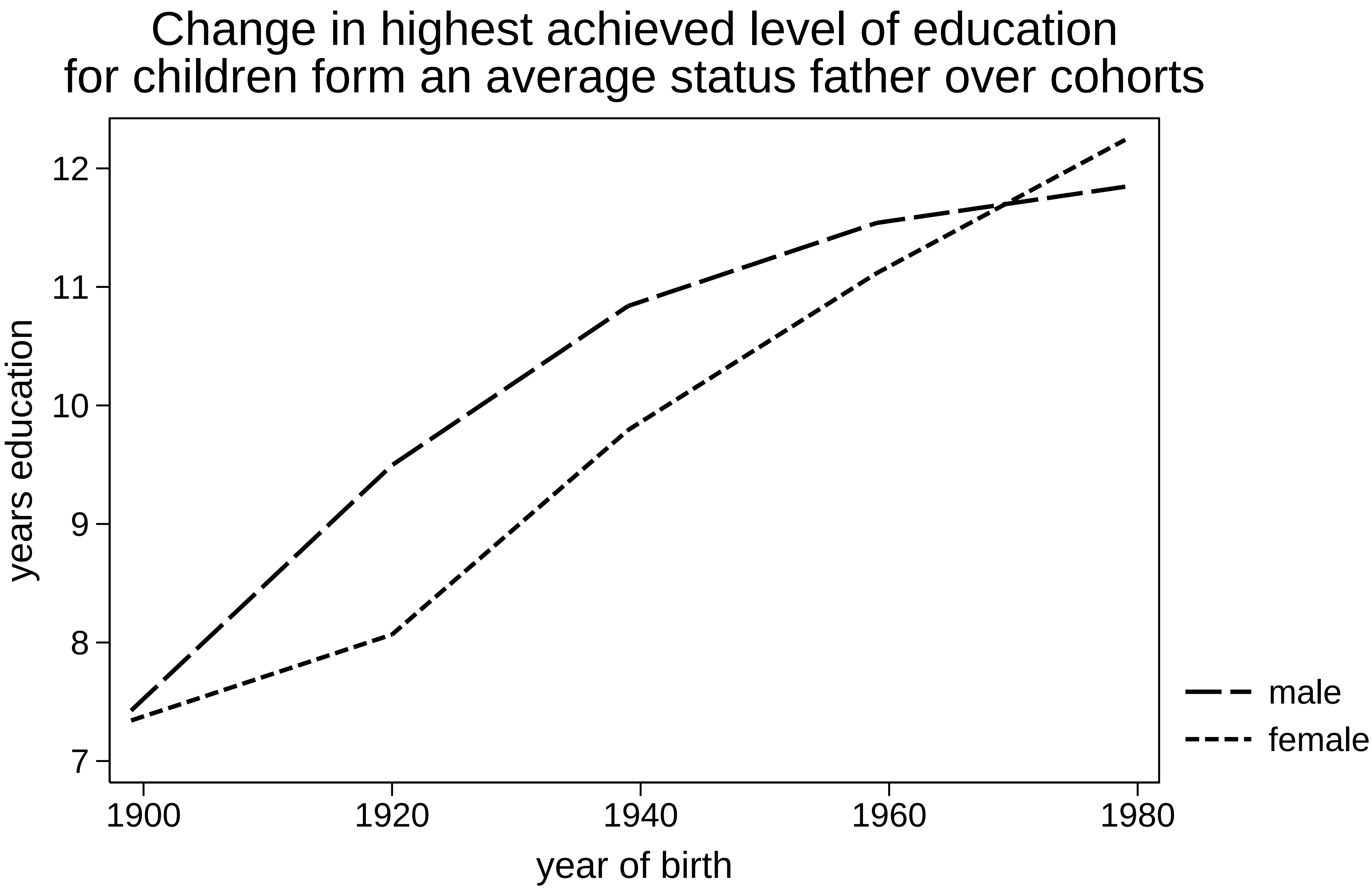
Baseline model

- Linear regression of highest achieved level of education on:
  - father's occupational status ($status$), which captures the Inequality of Educational Opportunity (IEO),
  - year of child's birth ($birthyear$), which captures educational expansion, and is added as a spline with three equally spaced knots to allow for non-linearity,
  - an interaction between $status$ and $birthyear$, which captures a linear trend in IEO,

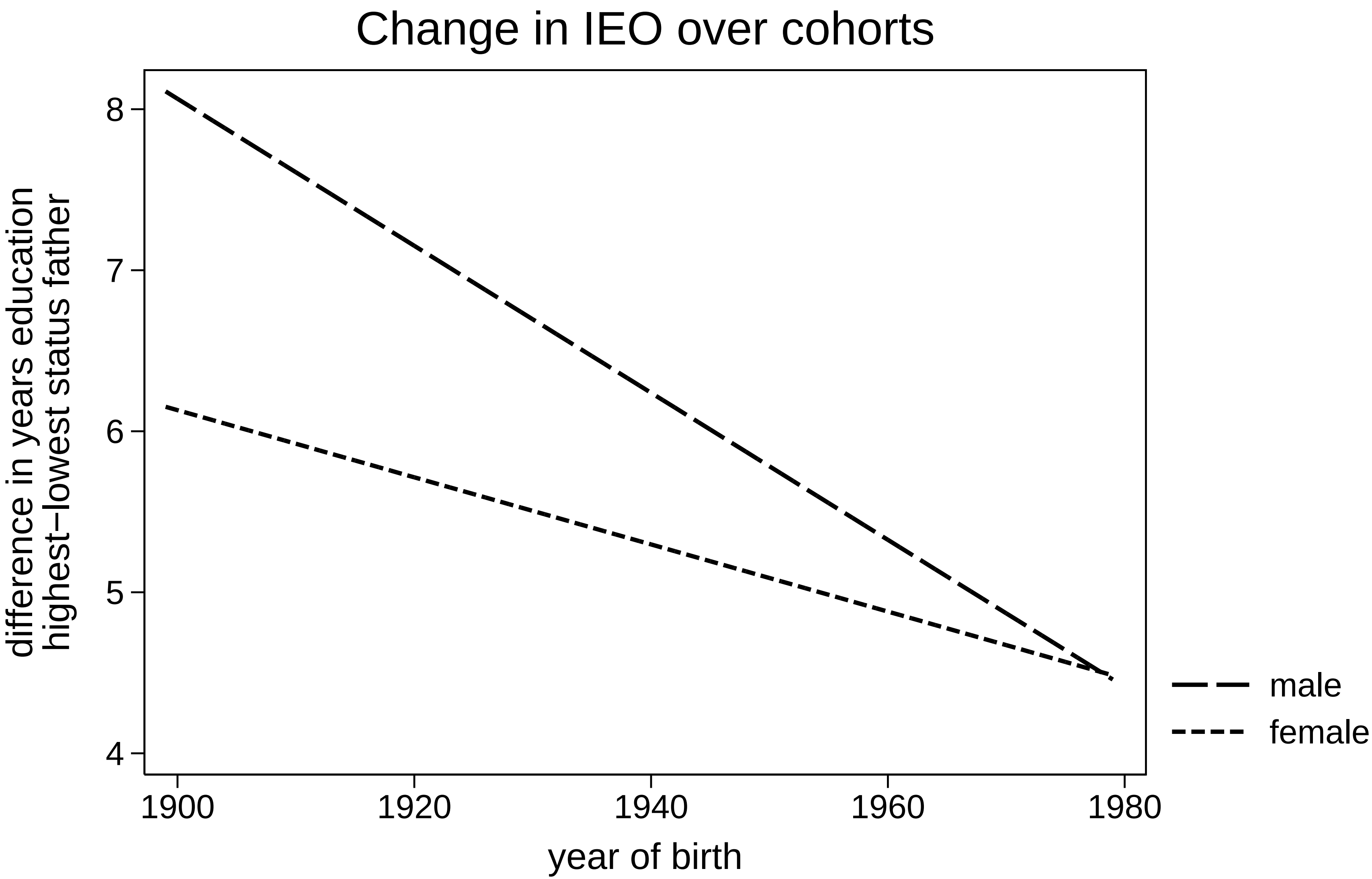
Baseline model

- Linear regression of highest achieved level of education on:
  - father's occupational status ($status$), which captures the Inequality of Educational Opportunity (IEO),
  - year of child's birth ($birthyear$), which captures educational expansion, and is added as a spline with three equally spaced knots to allow for non-linearity,
  - an interaction between $status$ and $birthyear$, which captures a linear trend in IEO,
  - and interactions of all variables with $female$.

Educational expansion in baseline model



IEO in baseline model



Outline

- baseline model
- Missing Data
  - Multiple Imputation of multiple surveys
  - assess plausibility of results
- Nesting within surveys
  - Random effects model
  - assess plausibility of results

Multiple Imputation

- Estimate for each missing value a distribution of plausible values.

Multiple Imputation

- Estimate for each missing value a distribution of plausible values.
- Draw multiple values from this distribution (typically 5), thus creating multiple ‘complete’ datasets.

Multiple Imputation

- Estimate for each missing value a distribution of plausible values.
- Draw multiple values from this distribution (typically 5), thus creating multiple ‘complete’ datasets.
- Estimate the model of interest on each ‘complete’ dataset.

Multiple Imputation

- Estimate for each missing value a distribution of plausible values.
- Draw multiple values from this distribution (typically 5), thus creating multiple ‘complete’ datasets.
- Estimate the model of interest on each ‘complete’ dataset.
- Point estimate is the average of the point estimates over the different ‘complete’ datasets.

Multiple Imputation

- Estimate for each missing value a distribution of plausible values.
- Draw multiple values from this distribution (typically 5), thus creating multiple ‘complete’ datasets.
- Estimate the model of interest on each ‘complete’ dataset.
- Point estimate is the average of the point estimates over the different ‘complete’ datasets.
- Variances of the point estimates are the averages of the variances in the different ‘complete’ datasets, plus a correction for the fact that the imputed cases weren’t real observations but only best guesses.

Multiple Imputation

- Estimate for each missing value a distribution of plausible values.
- Draw multiple values from this distribution (typically 5), thus creating multiple ‘complete’ datasets.
- Estimate the model of interest on each ‘complete’ dataset.
- Point estimate is the average of the point estimates over the different ‘complete’ datasets.
- Variances of the point estimates are the averages of the variances in the different ‘complete’ datasets, plus a correction for the fact that the imputed cases weren’t real observations but only best guesses.
- The correction is based on the between dataset variance of the point estimates.

Imputation model with multiple surveys

- The imputation model is a regression which must include at least all variables and interactions from the model of interest.

Imputation model with multiple surveys

- The imputation model is a regression which must include at least all variables and interactions from the model of interest.
- Separate models are estimated for each combination of survey, gender, and three year birthcohort to include all interactions and control for differences between surveys.

Imputation model with multiple surveys

- The imputation model is a regression which must include at least all variables and interactions from the model of interest.
- Separate models are estimated for each combination of survey, gender, and three year birthcohort to include all interactions and control for differences between surveys.
- Imputations are only made if enough complete observations are available (number of variables + 2).
  - Of 10,617 missing cases for $status$ 10,340 could be imputed.
  - Of 1,145 missing cases for $educyr$ 968 could be imputed.

Multiple Imputation results

		Complete Cases		Multiple Imputation	
		b	se	b	se
Male					
	$status$	8.065	0.252	8.038	0.252
	$birthyear X status$	-4.565	0.498	-4.554	0.500
Female					
	$status$	6.131	0.255	6.165	0.256
	$birthyear X status$	-2.085	0.493	-2.175	0.489

Multiple Imputation results

		Complete Cases		Multiple Imputation	
		b	se	b	se
Male					
	$status$	8.065	0.252	8.038	0.252
	$birthyearXstatus$	-4.565	0.498	-4.554	0.500
Female					
	$status$	6.131	0.255	6.165	0.256
	$birthyearXstatus$	-2.085	0.493	-2.175	0.489

Multiple Imputation results

		Complete Cases		Multiple Imputation	
		b	se	b	se
Male					
	$status$	8.065	0.252	8.038	0.252
	$birthyearXstatus$	-4.565	0.498	-4.554	0.500
Female					
	$status$	6.131	0.255	6.165	0.256
	$birthyearXstatus$	-2.085	0.493	-2.175	0.489

Diagnosing Imputation model

Asses the plausibility of results:

- How plausible is it that some standard errors in imputed model are larger than the standard errors in the complete case model?
- How plausible is it that the parameter estimates in the complete case model aren't biased?

Plausibility of increased SE

- With MI ‘new cases’ are added, so standard errors goes down, but not linearly.

Plausibility of increased SE

- With MI 'new cases' are added, so standard errors goes down, but not linearly.
- These 'new cases' are uncertain, and the correction for this uncertainty will make the standard error go up.

Plausibility of increased SE

- With MI 'new cases' are added, so standard errors goes down, but not linearly.
- These 'new cases' are uncertain, and the correction for this uncertainty will make the standard error go up.
- Standard error in regression does not only depend on N, but also on:
  - the standard deviation of the errors (fit of the model),
  - the correlation with other explanatory variables (multicollinearity), and
  - the variance of the explanatory variable itself.

Plausibility of increased SE

- With MI ‘new cases’ are added, so standard errors goes down, but not linearly.
- These ‘new cases’ are uncertain, and the correction for this uncertainty will make the standard error go up.
- Standard error in regression does not only depend on N, but also on:
  - the standard deviation of the errors (fit of the model),
  - the correlation with other explanatory variables (multicollinearity), and
  - the variance of the explanatory variable itself.
- Changes in these estimates may cause the standard error to go either up or down.

Decomposition of change in SE

Decomposition of change in SE relative to Complete Case SE

	sample size	imputation uncertainty	change in estimates†	total change
male				
$status$	-4.74%	0.25%	4.48%	-0.01%
$birthyear X status$	-4.74%	1.58%	3.46%	0.30%
female				
$status$	-4.74%	1.84%	3.35%	0.45%
$birthyear X status$	-4.74%	1.35%	2.58%	-0.81%

† standard deviation of the errors, degree of multicollinearity, and the variance of the explanatory variable

Plausibility of no bias

- Say we want to know $f(y|x)$, but x has missing values, so we know $f(y|x, M_x = 0)$.

Plausibility of no bias

- Say we want to know $f(y|x)$, but x has missing values, so we know $f(y|x, M_x = 0)$.
- Corrected estimates can be obtained by weighting the observations $\frac{\text{Pr}(M_x=0)}{\text{Pr}(M_x=0|y)}$.

Plausibility of no bias

- Say we want to know $f(y|x)$, but x has missing values, so we know $f(y|x, M_x = 0)$.
- Corrected estimates can be obtained by weighting the observations $\frac{\text{Pr}(M_x=0)}{\text{Pr}(M_x=0|y)}$.
- $\text{Pr}(M_x = 0)$ can be estimated by the proportion of complete observations.

Plausibility of no bias

- Say we want to know $f(y|x)$, but x has missing values, so we know $f(y|x, M_x = 0)$.
- Corrected estimates can be obtained by weighting the observations $\frac{\text{Pr}(M_x=0)}{\text{Pr}(M_x=0|y)}$.
- $\text{Pr}(M_x = 0)$ can be estimated by the proportion of complete observations.
- $\text{Pr}(M_x = 0|y)$ can be estimated using a logistic regression of M_x on y.

Weighting to correct for bias

$$f(y|x, M_x = 0) \quad = \quad \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)}$$

Weighting to correct for bias

$$f(y|x, M_x = 0) = \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)}$$

$$f(A|B, C) = \frac{f(A, B, C)}{f(B, C)}$$

Weighting to correct for bias

$$
\begin{aligned}
f(y|x, M_x = 0) &= \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)} \\
&= \frac{\textsf{Pr}(M_x = 0|y, x) f(y|x) f(x)}{\textsf{Pr}(M_x = 0|x) f(x)}
\end{aligned}
$$

Weighting to correct for bias

$$
\begin{aligned}
f(y|x, M_x = 0) &= \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)} \\
&= \frac{\mathsf{Pr}(M_x = 0|y, x) f(y|x) f(x)}{\mathsf{Pr}(M_x = 0|x) f(x)} \\
f(A, B, C) &= f(A|B, C) f(B|C) f(C)
\end{aligned}
$$

Weighting to correct for bias

$$\begin{aligned}
f(y|x, M_x = 0) &= \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)} \\
&= \frac{\textsf{Pr}(M_x = 0|y, x) f(y|x) f(x)}{\textsf{Pr}(M_x = 0|x) f(x)} \\
&= \frac{\textsf{Pr}(M_x = 0|y, x)}{\textsf{Pr}(M_x = 0|x)} f(y|x)
\end{aligned}$$

Weighting to correct for bias

$$
\begin{aligned}
f(y|x, M_x = 0) &= \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)} \\
&= \frac{\mathsf{Pr}(M_x = 0|y, x) f(y|x) f(x)}{\mathsf{Pr}(M_x = 0|x) f(x)} \\
&= \frac{\mathsf{Pr}(M_x = 0|y, x)}{\mathsf{Pr}(M_x = 0|x)} f(y|x) \\
&= \frac{\mathsf{Pr}(M_x = 0|y)}{\mathsf{Pr}(M_x = 0)} f(y|x) \quad \text{MAR assumption}
\end{aligned}
$$

Weighting to correct for bias

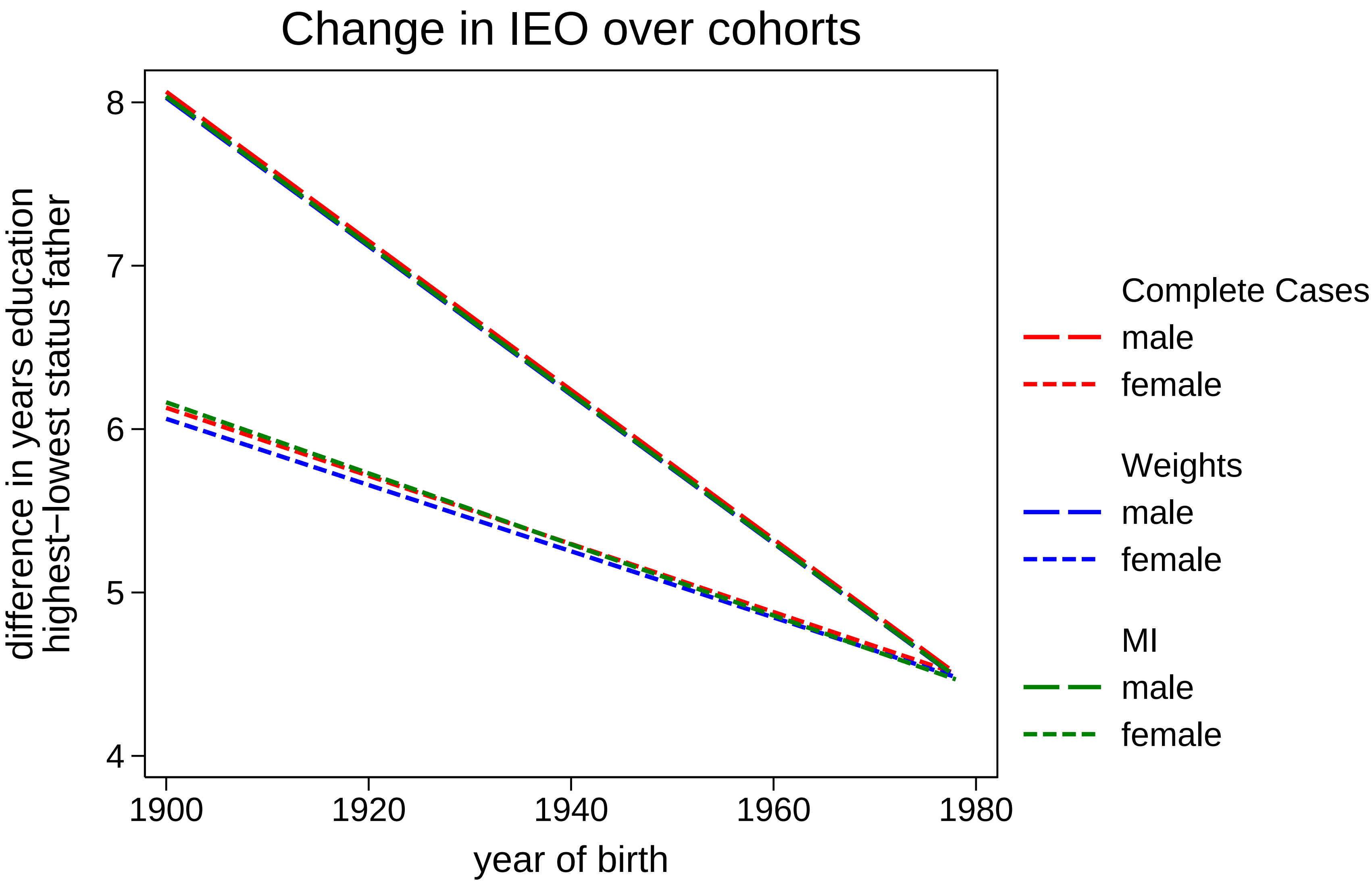
$$
\begin{aligned}
f(y|x, M_x = 0) &= \frac{f(y, x, M_x = 0)}{f(x, M_x = 0)} \\
&= \frac{\mathsf{Pr}(M_x = 0|y, x) f(y|x) f(x)}{\mathsf{Pr}(M_x = 0|x) f(x)} \\
&= \frac{\mathsf{Pr}(M_x = 0|y, x)}{\mathsf{Pr}(M_x = 0|x)} f(y|x) \\
&= \frac{\mathsf{Pr}(M_x = 0|y)}{\mathsf{Pr}(M_x = 0)} f(y|x) \quad \text{MAR assumption} \\
f(y|x) &= \frac{\mathsf{Pr}(M_x = 0)}{\mathsf{Pr}(M_x = 0|y)} f(y|x, M_x = 0)
\end{aligned}
$$

Weighting to correct for bias

This approach can be extended to include:

- missing cases in y,
- multiple xs with or without missing cases,
- interaction terms.

IEO with corrections for missing data



Outline

- baseline model
- Missing Data
  - Multiple Imputation of multiple surveys
  - assess plausibility of results
- Nesting within surveys
  - Random effects model
  - assess plausibility of results

Nested structure of the data

Random effects model:

- Random effects:
  - (Male) constant
  - (Male) $status$
- Fixed effects:
  - $female$
  - $femaleXstatus$
  - splines of $birthyear$
  - trends in $status$

Random effects model

		Pooled regression		Random effects	
		b	se	b	se
Male					
	$status$	8.065	0.252	8.469	0.297
	$birthyear X status$	-4.565	0.498	-5.542	0.549
Female					
	$status$	6.131	0.255	6.636	0.298
	$birthyear X status$	-2.085	0.493	-3.305	0.543

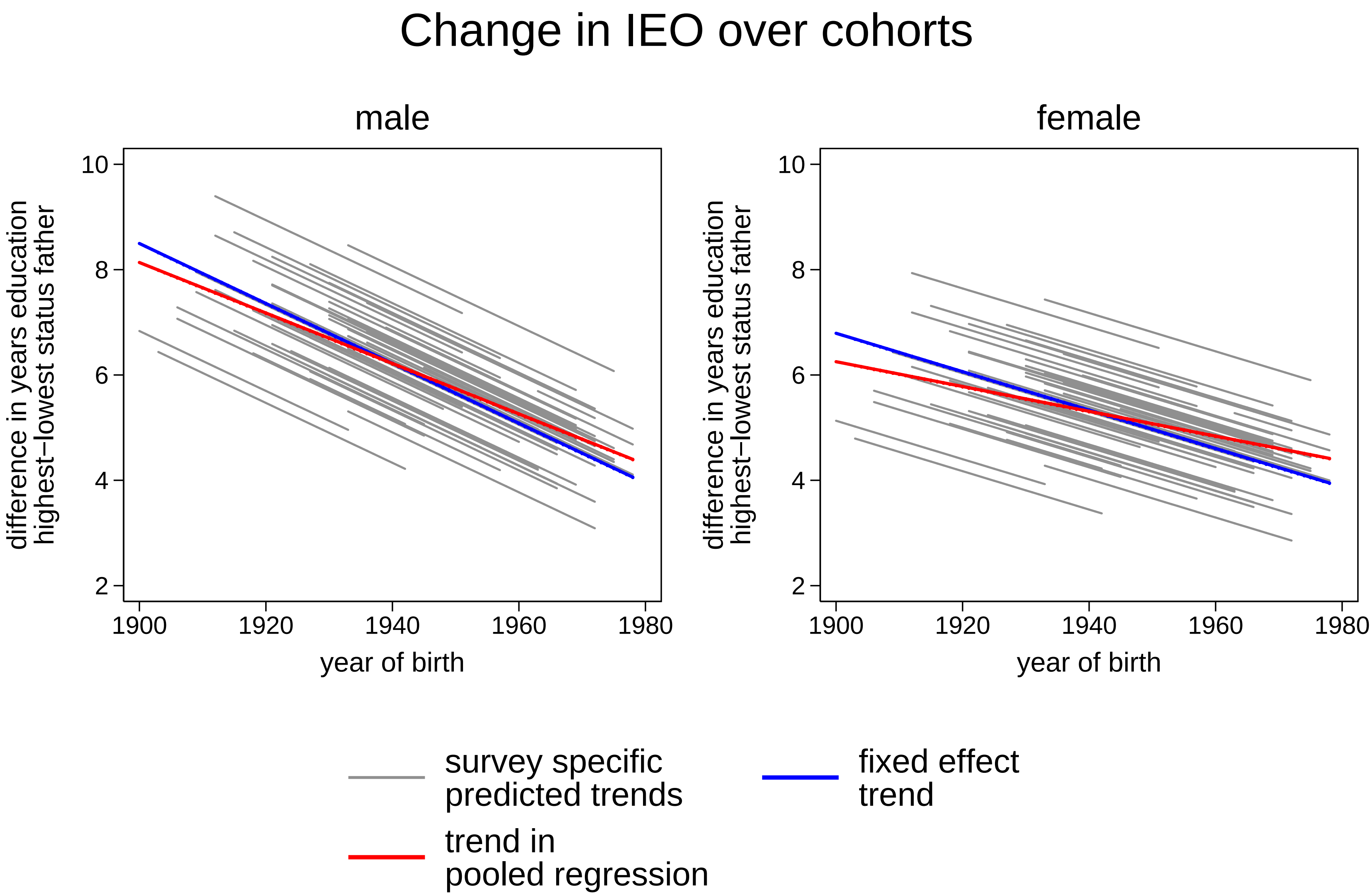
Random effects model

		Pooled regression		Random effects	
		b	se	b	se
Male					
	$status$	8.065	0.252	8.469	0.297
	$birthyear X status$	-4.565	0.498	-5.542	0.549
Female					
	$status$	6.131	0.255	6.636	0.298
	$birthyear X status$	-2.085	0.493	-3.305	0.543

Random effects model

		Pooled regression		Random effects	
		b	se	b	se
Male					
	$status$	8.065	0.252	8.469	0.297
	$birthyear X status$	-4.565	0.498	-5.542	0.549
Female					
	$status$	6.131	0.255	6.636	0.298
	$birthyear X status$	-2.085	0.493	-3.305	0.543

Plausibility of bias in Pooled regression



Outlying surveys

- Three outlying surveys:
  - Gadourek 1958, ‘Health threatening habits’,
  - Kooij 1967, ‘Family in modern city environment’, and
  - ISSP 1999, ‘Social Inequality III’.

Outlying surveys

- Three outlying surveys:
  - Gadourek 1958, ‘Health threatening habits’,
  - Kooij 1967, ‘Family in modern city environment’, and
  - ISSP 1999, ‘Social Inequality III’.
- The level of IEO is either underestimated (early surveys) or overestimated (late surveys), so in a pooled regression these lead to an underestimation of the trend in IEO.

Outlying surveys

- Three outlying surveys:
  - Gadourek 1958, ‘Health threatening habits’,
  - Kooij 1967, ‘Family in modern city environment’, and
  - ISSP 1999, ‘Social Inequality III’.
- The level of IEO is either underestimated (early surveys) or overestimated (late surveys), so in a pooled regression these lead to an underestimation of the trend in IEO.
- Trend in inequality within surveys is pretty consistent.

Outlying surveys

- Three outlying surveys:
  - Gadourek 1958, ‘Health threatening habits’,
  - Kooij 1967, ‘Family in modern city environment’, and
  - ISSP 1999, ‘Social Inequality III’.
- The level of IEO is either underestimated (early surveys) or overestimated (late surveys), so in a pooled regression these lead to an underestimation of the trend in IEO.
- Trend in inequality within surveys is pretty consistent.
- These studies provide valuable information about the trend once one controls for level of IEO.

Conclusions

- Missing data
  - Virtually no bias was found.
  - Virtually no gain in power was achieved by using Multiple Imputation.
- Nested structure of the data
  - Outlying studies have lead to an underestimation of the trend in IEO in pooled regression.
  - Standard errors increases a little when controlling for nested structure.