Assessing the reasonableness of an imputation model

Maarten L. Buis

Department of Social Research Methodology
Vrije Universiteit Amsterdam
http://home.fsw.vu.nl/m.buis/

Outline

Outline

- Missing Data
- Multiple Imputation
- Weighting
  - theory
  - weightmis
- Application

Missing data

- two problems:
  1. Loss of information
  2. bias

Missing data

- two problems:
    1. Loss of information
    2. bias
- Solution: Multiple Imputation

Missing data

- two problems:
  1. Loss of information
  2. bias
- Solution: Multiple Imputation
- model diagnostics:
  - Plot distribution of observed and imputed values (Royston 2005a, Abayomi, Gelman, Levy 2006)

Missing data

- two problems:
  1. Loss of information
  2. bias
- Solution: Multiple Imputation
- model diagnostics:
  - Plot distribution of observed and imputed values (Royston 2005a, Abayomi, Gelman, Levy 2006)
  - Check whether imputation algorithm has converged (Royston 2005b)

Missing data

- two problems:
  1. Loss of information
  2. bias
- Solution: Multiple Imputation
- model diagnostics:
  - Plot distribution of observed and imputed values (Royston 2005a, Abayomi, Gelman, Levy 2006)
  - Check whether imputation algorithm has converged (Royston 2005b)
  - compare results with alternative method

Missing data

- two problems:
  1. Loss of information
  2. bias
- Solution: Multiple Imputation
- model diagnostics:
  - Plot distribution of observed and imputed values (Royston 2005a, Abayomi, Gelman, Levy 2006)
  - Check whether imputation algorithm has converged (Royston 2005b)
  - compare results with alternative method: weighting

Three types missingness

1. Missing Completely At Random (MCAR)
    - Probability of being missing does not depend on any other variable.
    - Complete data is a random subsample of the original sample. So, loss of information, but no bias.

Three types missingness

1. Missing Completely At Random (MCAR)
   - Probability of being missing does not depend on any other variable.
   - Complete data is a random subsample of the original sample. So, loss of information, but no bias.
2. Missing At Random (MAR)
   - Probability of being missing depends on other variables but not on the missing value itself.
   - Both potential bias and loss of information.

Three types missingness

1. Missing Completely At Random (MCAR)
    - Probability of being missing does not depend on any other variable.
    - Complete data is a random subsample of the original sample. So, loss of information, but no bias.
2. Missing At Random (MAR)
    - Probability of being missing depends on other variables but not on the missing value itself.
    - Both potential bias and loss of information.
3. Not Missing At Random (NMAR)
    - Probability of being missing depends on the missing value itself.
    - Both potential bias and loss of information.

Outline

Missing Data

Multiple Imputation

Weighting
- theory
- weightmis

Application

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.

Multiple Imputation in Stata

- Within Stata the distribution of plausible values can be estimated with `ice` and `hotdeck`.
- Within Stata the estimates from the ‘complete’ datasets can be combined with `mim`.

Outline

Missing Data

Multiple Imputation

Weighting
- theory
- weightmis

Application

Missing values for one x.

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

Missing values for one x.

Bayes' Rule

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

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

Missing values for one x.

Bayes' Rule

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

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

Missing values for one x.

Bayes' Rule

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

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

Missing values for one x.

Bayes' Rule

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

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

Missing values for one x.

$$f(y|x, R_x) = \frac{f(y, x, R_x)}{f(x, R_x)} = \frac{\Pr(R_x|y, x)f(y|x)f(x)}{\Pr(R_x|x)f(x)}$$

Missing values for one x.

Bayes' Rule again

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

Missing values for one x.

Bayes' Rule again

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

Missing values for one x.

Bayes' Rule again

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

Missing values for one x.

Bayes' Rule again

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

Missing values for one x.

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

Missing values for one x.

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

Missing values for one x.

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

Missing values for one x.

MAR assumption

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

Missing values for one x.

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

Estimating the weights $\frac{\Pr(R_x)}{\Pr(R_x|y)}$

1. Create a variable indicating whether or not x is observed:

```
gen Rx = !missing(x)
```

Estimating the weights $\frac{\Pr(R_x)}{\Pr(R_x|y)}$

1. Create a variable indicating whether or not *x* is observed:
```
gen Rx = !missing(x)
```
2. Estimate $\Pr(R_x)$ by:
```
logit Rx
predict PrRx, pr
```

Estimating the weights $\frac{\Pr(R_x)}{\Pr(R_x|y)}$

1. Create a variable indicating whether or not *x* is observed:
```
gen Rx = !missing(x)
```
2. Estimate $\Pr(R_x)$ by:
```
logit Rx
predict PrRx, pr
```
3. Estimate $\Pr(R_x|y)$ by:
```
logit Rx y
predict PrRxGy, pr
```

Estimating the weights $\frac{\Pr(R_x)}{\Pr(R_x|y)}$

1. Create a variable indicating whether or not x is observed:
```
gen Rx = !missing(x)
```
2. Estimate $\Pr(R_x)$ by:
```
logit Rx
predict PrRx, pr
```
3. Estimate $\Pr(R_x|y)$ by:
```
logit Rx y
predict PrRxGy, pr
```
4. generate the weight by:
```
gen w = PrRx/PrRxGy
```

Missing values for two *x*s and *y*.

Bayes' Rule

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y, x_1, x_2, R_{x_1}, R_{x_2}, R_y)}{f(x_1, x_2, R_{x1}, R_{x2}, R_y)}$$

Missing values for two *x*s and *y*.

Bayes' Rule again

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

Missing values for two *x*s and *y*.

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y, x_1, x_2, R_{x_1}, R_{x_2}, R_y)}{f(x_1, x_2, R_{x1}, R_{x2}, R_y)}$$

$$= \frac{\Pr(R_{x_1}|y, x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|y, x_1, x_2, R_y)\Pr(R_y|y, x_1, x_2)f(y|x_1, x_2)f(x_1, x_2)}{\Pr(R_{x_1}|x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|x_1, x_2, R_y)\Pr(R_y|x_1, x_2)f(x_1, x_2)}$$

$$= \frac{\Pr(R_{x_1}|y, x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|y, x_1, x_2, R_y)\Pr(R_y|y, x_1, x_2)}{\Pr(R_{x_1}|x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|x_1, x_2, R_y)\Pr(R_y|x_1, x_2)} f(y|x_1, x_2)$$

Missing values for two *x*s and *y*.

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y, x_1, x_2, R_{x_1}, R_{x_2}, R_y)}{f(x_1, x_2, R_{x1}, R_{x2}, R_y)}$$

$$= \frac{\Pr(R_{x_1}|y, x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|y, x_1, x_2, R_y)\Pr(R_y|y, x_1, x_2)f(y|x_1, x_2)f(x_1, x_2)}{\Pr(R_{x_1}|x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|x_1, x_2, R_y)\Pr(R_y|x_1, x_2)f(x_1, x_2)}$$

$$= \frac{\Pr(R_{x_1}|y, x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|y, x_1, x_2, R_y)\Pr(R_y|y, x_1, x_2)}{\Pr(R_{x_1}|x_1, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|x_1, x_2, R_y)\Pr(R_y|x_1, x_2)} f(y|x_1, x_2)$$

Missing values for two xs and y.

MAR assumption

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

$$= \frac{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)\Pr(R_y|x_1,x_2)}{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

Missing values for two *x*s and *y*.

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

$$= \frac{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)\Pr(R_y|x_1,x_2)}{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

Missing values for two *x*s and *y*.

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

$$= \frac{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)} f(y|x_1, x_2)$$

Missing values for two *x*s and *y*.

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

$$= \frac{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)} f(y|x_1, x_2)$$

$$f(y|x_1, x_2) = \frac{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)}{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)} f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y)$$

Missing values for two xs and y.

Observed

$$f(y|x_1,x_2,R_{x_1},R_{x_2},R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1,x_2)$$

$$= \frac{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)} f(y|x_1,x_2)$$

$$f(y|x_1,x_2) = \frac{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)}{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)} f(y|x_1,x_2,R_{x_1},R_{x_2},R_y)$$

Missing values for two xs and y.

Not observed if x_1 is missing

$$f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y) = \frac{f(y,x_1,x_2,R_{x_1},R_{x_2},R_y)}{f(x_1,x_2,R_{x1},R_{x2},R_y)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)f(y|x_1,x_2)f(x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)f(x_1,x_2)}$$

$$= \frac{\Pr(R_{x_1}|y,x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,x_2,R_y)\Pr(R_y|y,x_1,x_2)}{\Pr(R_{x_1}|x_1,x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,x_2,R_y)\Pr(R_y|x_1,x_2)} f(y|x_1, x_2)$$

$$= \frac{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)} f(y|x_1, x_2)$$

$$f(y|x_1, x_2) = \frac{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)}{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)} f(y|x_1, x_2, R_{x_1}, R_{x_2}, R_y)$$

Estimating the weight $\frac{\Pr(R_{x_1}|x_2, R_{x_2}, R_y)\Pr(R_{x_2}|x_1, R_y)}{\Pr(R_{x_1}|y, x_2, R_{x_2}, R_y)\Pr(R_{x_2}|y, x_1, R_y)}$

1. The weight can be split up into two parts:

$$\frac{\Pr(R_{x_1}|x_2, R_{x_2}, R_y)}{\Pr(R_{x_1}|y, x_2, R_{x_2}, R_y)} \times \frac{\Pr(R_{x_2}|x_1, R_y)}{\Pr(R_{x_2}|y, x_1, R_y)}$$

Estimating the weight $\frac{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)}{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}$

1. The weight can be split up into two parts:

$$\frac{\Pr(R_{x_1}|x_2, R_{x_2}, R_y)}{\Pr(R_{x_1}|y, x_2, R_{x_2}, R_y)} \times \frac{\Pr(R_{x_2}|x_1, R_y)}{\Pr(R_{x_2}|y, x_1, R_y)}$$

2. For both the first and the second part only use cases which are observed on y.

Estimating the weight $\frac{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)}{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}$

1. The weight can be split up into two parts:

$$\frac{\Pr(R_{x_1}|x_2, R_{x_2}, R_y)}{\Pr(R_{x_1}|y, x_2, R_{x_2}, R_y)} \times \frac{\Pr(R_{x_2}|x_1, R_y)}{\Pr(R_{x_2}|y, x_1, R_y)}$$

2. For both the first and the second part only use cases which are observed on y.
3. The first part can be estimated like before with `logit` and `predict`.

Estimating the weight $\frac{\Pr(R_{x_1}|x_2,R_{x_2},R_y)\Pr(R_{x_2}|x_1,R_y)}{\Pr(R_{x_1}|y,x_2,R_{x_2},R_y)\Pr(R_{x_2}|y,x_1,R_y)}$

1. The weight can be split up into two parts:

$$\frac{\Pr(R_{x_1}|x_2, R_{x_2}, R_y)}{\Pr(R_{x_1}|y, x_2, R_{x_2}, R_y)} \times \frac{\Pr(R_{x_2}|x_1, R_y)}{\Pr(R_{x_2}|y, x_1, R_y)}$$

2. For both the first and the second part only use cases which are observed on y.
3. The first part can be estimated like before with `logit` and `predict`.
4. The second part can be estimated with `logit` and `predict`, but now with weights to correct for missing data in x_1.

A recursive algorithm

- In other words: With two *x*s with missing data the algorithm calls itself twice to solve two smaller missing data problems.

A recursive algorithm

- In other words: With two *x*s with missing data the algorithm calls itself twice to solve two smaller missing data problems.
- In principle this method could be expanded for any number of *x*s with missing data,

A recursive algorithm

- In other words: With two *x*s with missing data the algorithm calls itself twice to solve two smaller missing data problems.
- In principle this method could be expanded for any number of *x*s with missing data,
- but the number of calls to `logit` rises very quickly with the number of variables.

number of variables	1	2	3	4	5	6
number of calls to `logit`	2	8	22	52	114	240

Number of variables

- Often the same variable enters a regression equation multiple time, e.g.:
  - interaction terms
  - dummy variables
  - polynomials
  - splines

Number of variables

- Often the same variable enters a regression equation multiple time, e.g.:
  - interaction terms
  - dummy variables
  - polynomials
  - splines
- These variables count as one variable, thus diminishing the computational load.

weightmis syntax

`weightmis` *varlist* [*if*] [*in*] [*pw*], command(*string*) [missing(*varlist*) observed(*varlist*) double#(*varlist*) generate(*string*) *]

example 1

Say, y, x_1, and x_2 contain missing values, and you want to estimate the following regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

```
weightmis y x1 x2, command(regress) /*
*/ missing(x1 x2)
```

example 2

Say, y, x_1, and x_2 contain missing values, and you want to estimate the following regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2 + \varepsilon$$

```
weightmis y x1 x2 x2sq, command(regress) /*
/* missing(x1 x2) double2(x2sq)
```

example 3

Say, y, x_1, and x_2 contain missing values, and you want to estimate the following regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

```
weightmis y x1 x2 x1x2, command(regress) /*
*/ missing(x1 x2) double1(x1x2) double2(x1x2)
```

Outline

Data

- The aim is to look at the strength of association between family background and child's highest achieved level of education

Data

- The aim is to look at the strength of association between family background and child's highest achieved level of education, inequality of educational opportunity.

Data

- The aim is to look at the strength of association between family background and child's highest achieved level of education, inequality of educational opportunity.
- International Stratification and Mobility File (ISMF) on the Netherlands.

Data

- The aim is to look at the strength of association between family background and child's highest achieved level of education, inequality of educational opportunity.
- International Stratification and Mobility File (ISMF) on the Netherlands.
- 51 surveys held between 1958 and 2005 with information on cohorts 1906-1990.

Data

- The aim is to look at the strength of association between family background and child's highest achieved level of education, inequality of educational opportunity.
- International Stratification and Mobility File (ISMF) on the Netherlands.
- 51 surveys held between 1958 and 2005 with information on cohorts 1906-1990.
- 96,761 respondents aged between 27 and 65.

Data

- The aim is to look at the strength of association between family background and child's highest achieved level of education, inequality of educational opportunity.
- International Stratification and Mobility File (ISMF) on the Netherlands.
- 51 surveys held between 1958 and 2005 with information on cohorts 1906-1990.
- 96,761 respondents aged between 27 and 65.
- Number of cases are unequally distributed over cohorts.

Model

- Linear regression of highest achieved level of education (*educyr*) on:
  - father's occupational status (*fisei*),

Model

- Linear regression of highest achieved level of education (*educyr*) on:
    - father's occupational status (*fisei*),
    - Year in which the child is 12 (*byr*), and is added as a spline with three knots to allow for non-linearity,

Model

- Linear regression of highest achieved level of education (*educyr*) on:
  - father's occupational status (*fisei*),
  - Year in which the child is 12 (*byr*), and is added as a spline with three knots to allow for non-linearity,
  - an interaction between *fisei* and the splines of *byr*,

Model

- Linear regression of highest achieved level of education (*educyr*) on:
  - father's occupational status (*fisei*),
  - Year in which the child is 12 (*byr*), and is added as a spline with three knots to allow for non-linearity,
  - an interaction between *fisei* and the splines of *byr*,
  - and interactions of all variables with *female*.

Summary of missing values using misschk

```
   #  Variable         # Missing   % Missing
--------------------------------------------
   1  educyr              1125         1.2
   2  fisei              10082        10.4
   3  female                 0         0.0
   4  byr                    0         0.0

Missing for |
      which |
 variables? |      Freq.     Percent        Cum.
------------+-----------------------------------
       12__ |        330        0.34        0.34
       1___ |        795        0.82        1.16
       _2__ |      9,752       10.08       11.24
       ____ |     85,884       88.76      100.00
------------+-----------------------------------
      Total |     96,761      100.00
```

Imputation model

- Regress *fisei* on *educyr*, *female*, *byr* (in dummies), dummies for survey, and all interactions.

Imputation model

- Regress *fisei* on *educyr*, *female*, *byr* (in dummies), dummies for survey, and all interactions.
- For each missing value of *fisei* draw a random value from a normal distribution whose mean is the predicted value of *fisei* and and whose standard deviation is the standard deviation of the errors.

Imputation model

- Regress *fisei* on *educyr*, *female*, *byr* (in dummies), dummies for survey, and all interactions.
- For each missing value of *fisei* draw a random value from a normal distribution whose mean is the predicted value of *fisei* and and whose standard deviation is the standard deviation of the errors.
- Predictions can be improved by adding other variables, like father's education (*feducyr*), mother's education(*meducyr*), child's occupational status (*isei*).

Imputation model

- In practice the interactions with survey number, *female*, and *byr* are modeled by estimating separate models for each combination of survey, gender, and three year birthcohort.

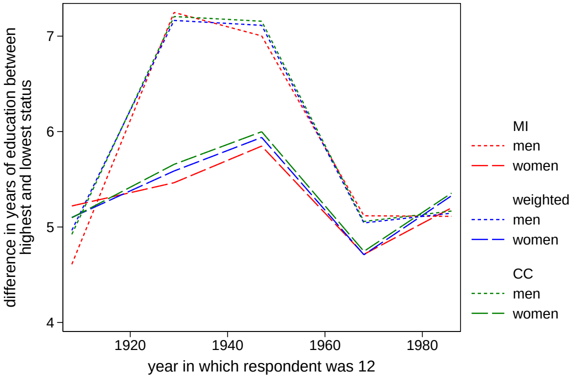
Imputation model

- In practice the interactions with survey number, *female*, and *byr* are modeled by estimating separate models for each combination of survey, gender, and three year birthcohort.
- *feducyr*, and *meducyr* are only used if they were asked in that survey.

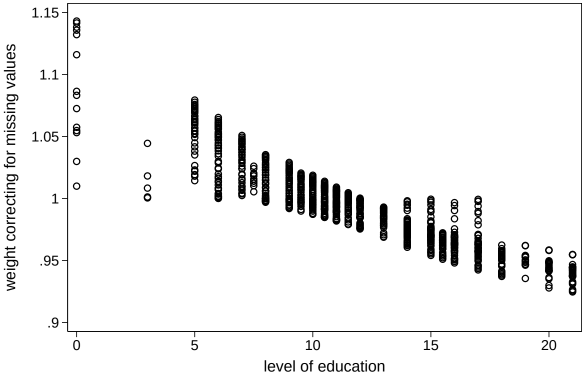
Imputation model

- In practice the interactions with survey number, *female*, and *byr* are modeled by estimating separate models for each combination of survey, gender, and three year birthcohort.
- *feducyr*, and *meducyr* are only used if they were asked in that survey.
- Imputations are only made if enough complete observations are available (number of variables + 2).
  - Of 10,082 missing cases for *fisei* 191 could not be imputed.
  - Of 1,145 missing cases for *educyr* 148 could not be imputed.

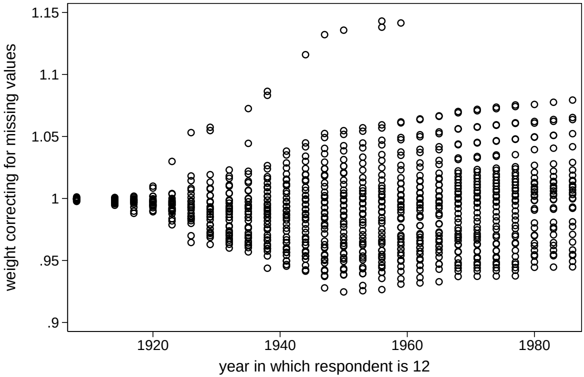
Trends in Inequality of educational opportunity



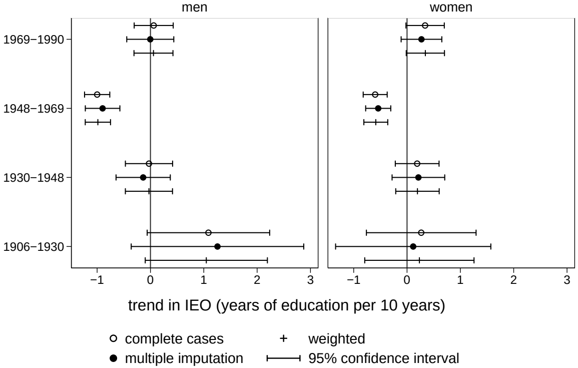
Weight versus level of education



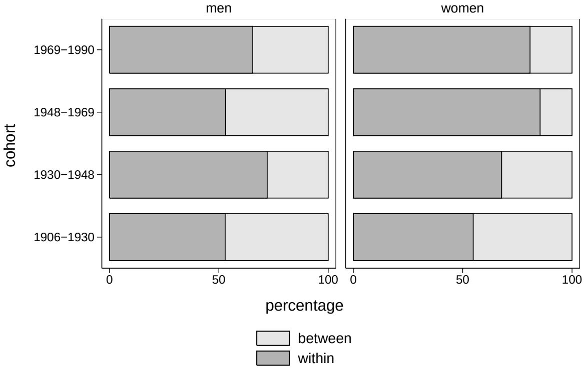
Weight versus cohort



Confidence intervals



Percentage of variance due to average variance across datasets and variance between datasets



Conclusion

- The imputation model becomes part of the statistical model when using Multiple Imputation, and needs to be checked.

Conclusion

- The imputation model becomes part of the statistical model when using Multiple Imputation, and needs to be checked.
- One possible way of doing that is to compare the results with an alternative method that should also result in valid results.

Conclusion

- The imputation model becomes part of the statistical model when using Multiple Imputation, and needs to be checked.
- One possible way of doing that is to compare the results with an alternative method that should also result in valid results.
- One such method is weighting, as (to be) implemented in `weightmis`

References

Patrick Royston.
Multiple Imputation of Missing Values: Update.
The Stata Journal, 5(2):188–201, 2005a.

Patrick Royston.
Multiple Imputation of Missing Values: Update of ice.
The Stata Journal, 5(4):527–636, 2005b.

Kobi Abayomi, Andrew Gelman, Marc Levy.
Diagnostics for Multivariate Imputations.
`http://www.stat.columbia.edu/~gelman/research/unpublished/paper73.pdf` 2006