INTERMEDIATE QUANTITATIVE METHODS

Measurement Error and Missing Data

Per Engzell (Nuffield College)

Hilary term 2018

Sociology Department, University of Oxford

MISSING DATA

MISSING DATA

When observations that should be in the sample are missing

The most common reason for this in surveys is nonresponse

- Some respondents cannot be reached or refuse to answer (”unit nonresponse”)
- Others fail to answer specific questions (” item nonresponse”)

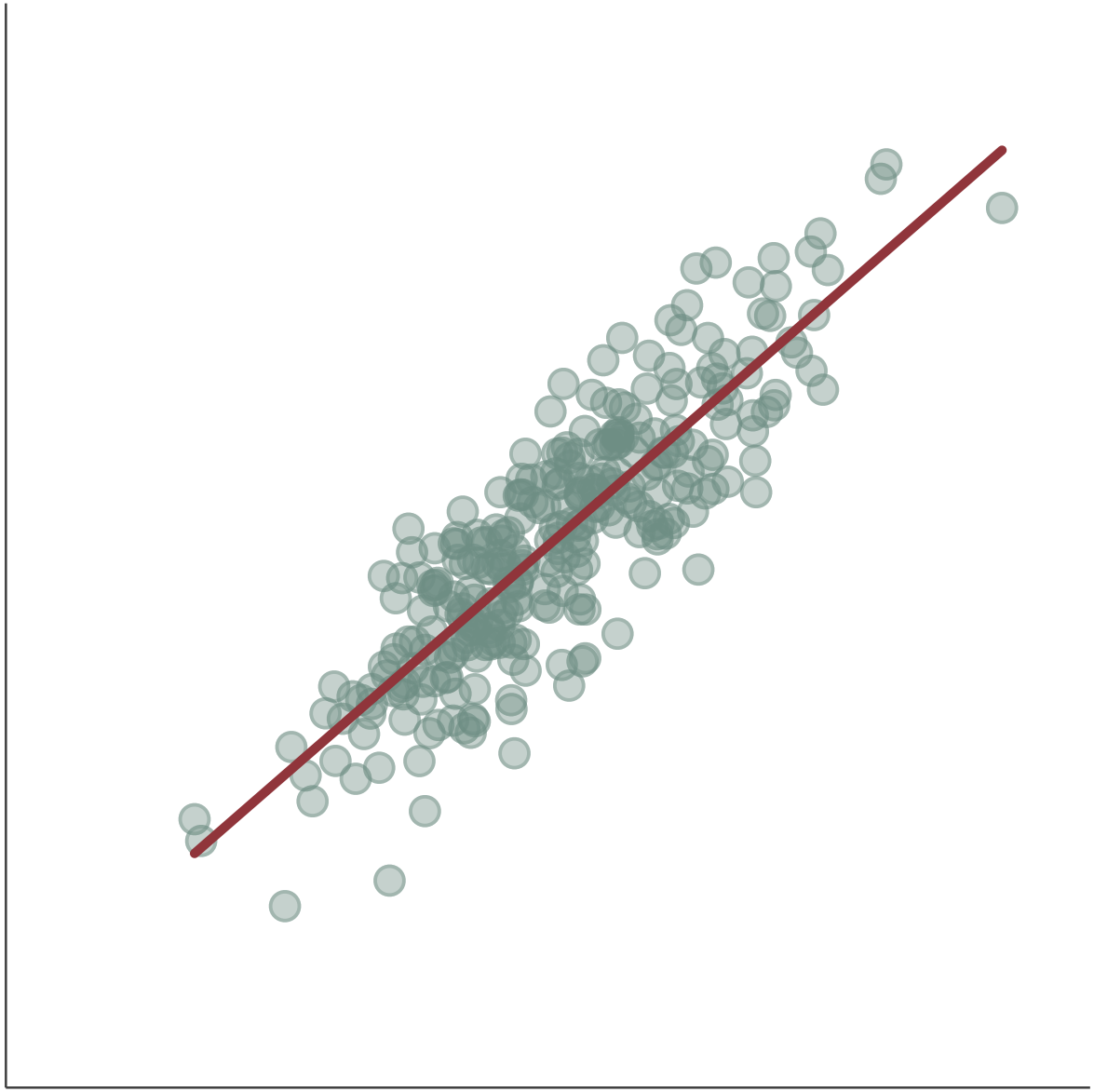
The worry is that those who are lacking answers are systematically different – nonresponse is selective

How does this affect the conclusions we are able to draw?

UNIT NONRESPONSE

Selection on y

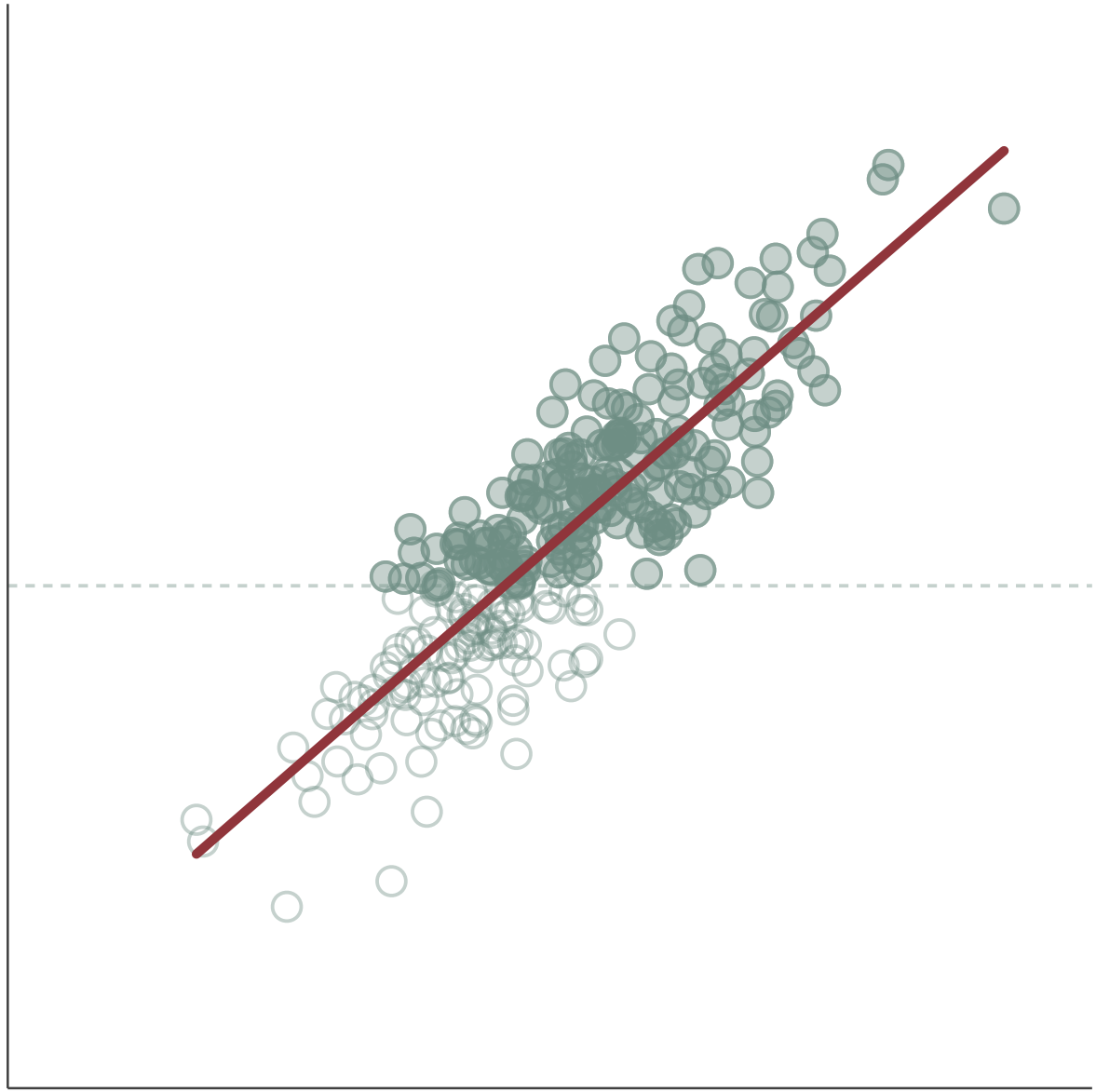
- True population relationship



UNIT NONRESPONSE

Selection on y

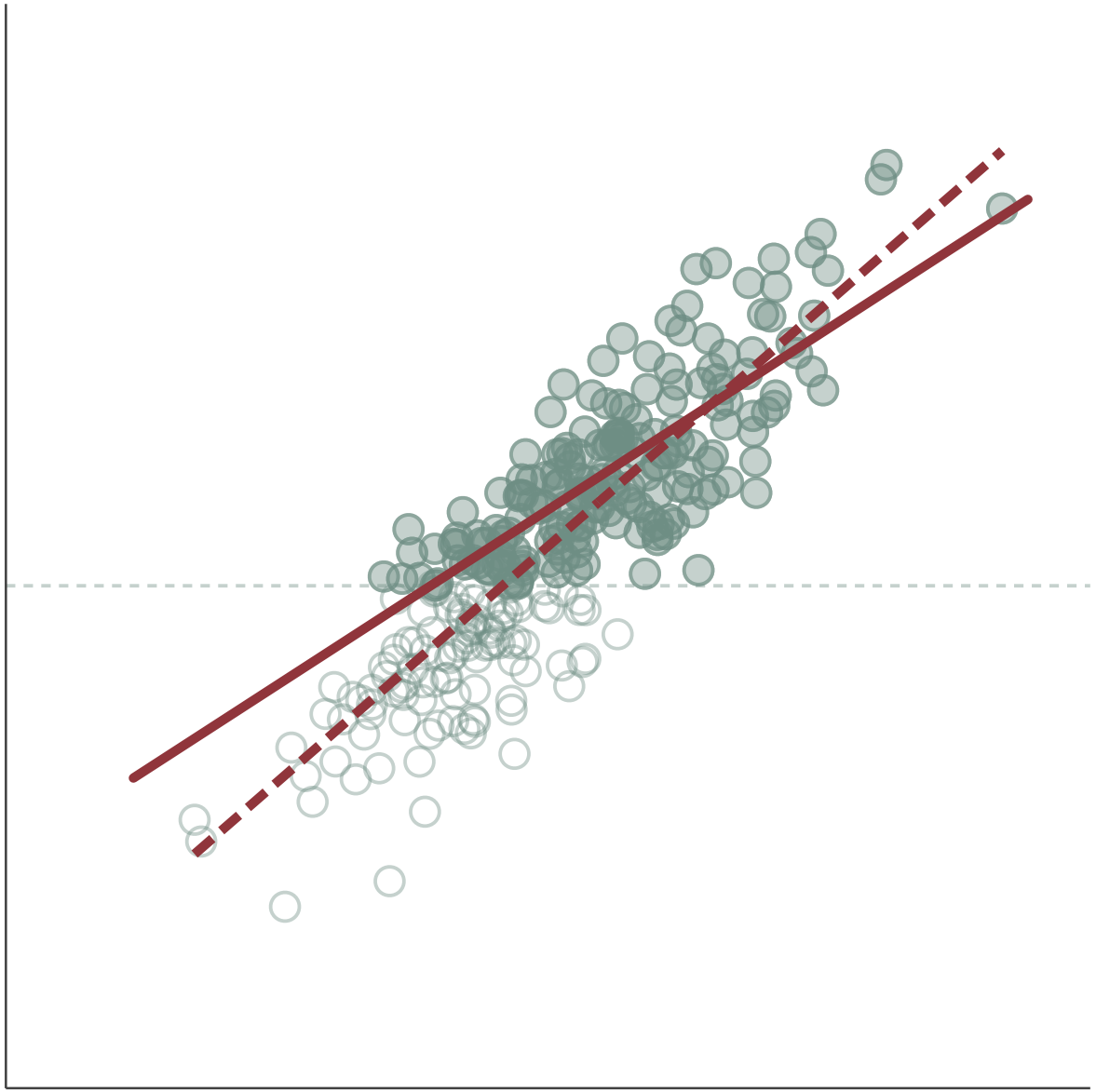
- True population relationship
- Estimated only observing those with high y



UNIT NONRESPONSE

Selection on y

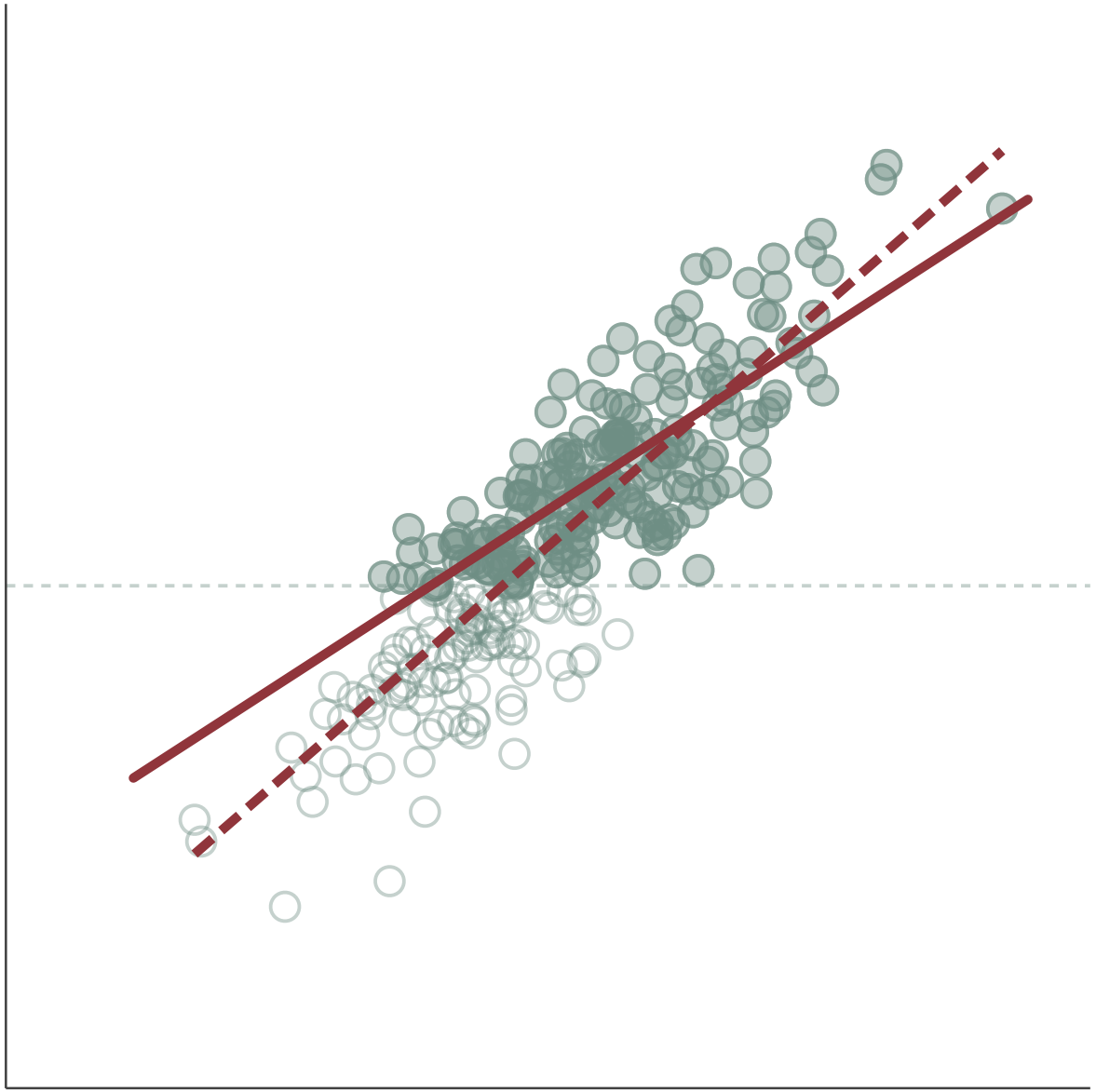
- True population relationship
- Estimated only observing those with high y
- Estimated relationship is weaker



UNIT NONRESPONSE

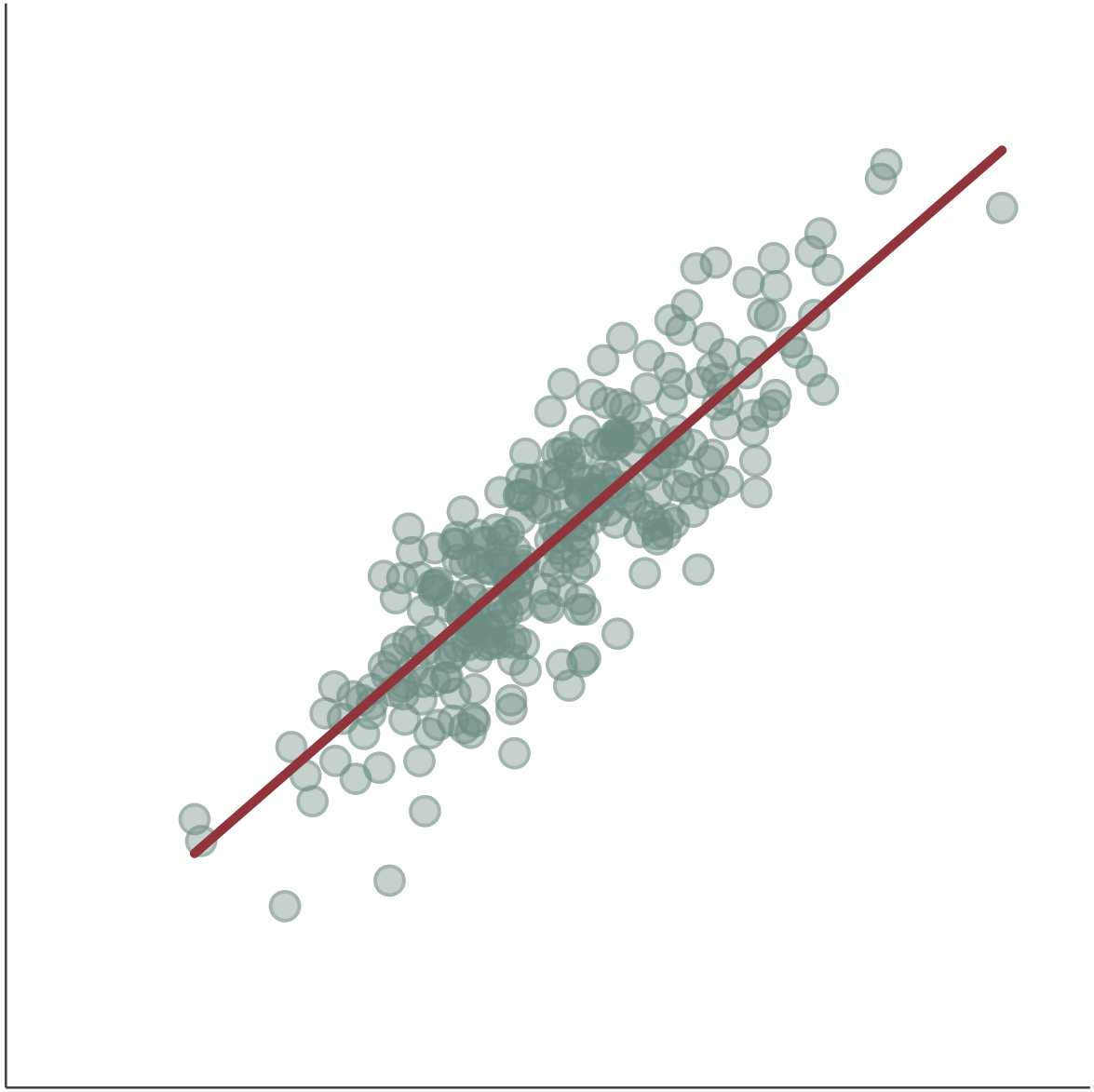
Selection on y

- True population relationship
- Estimated only observing those with high y
- Estimated relationship is weaker
- Also: smaller n, larger sampling variance



Selection on x

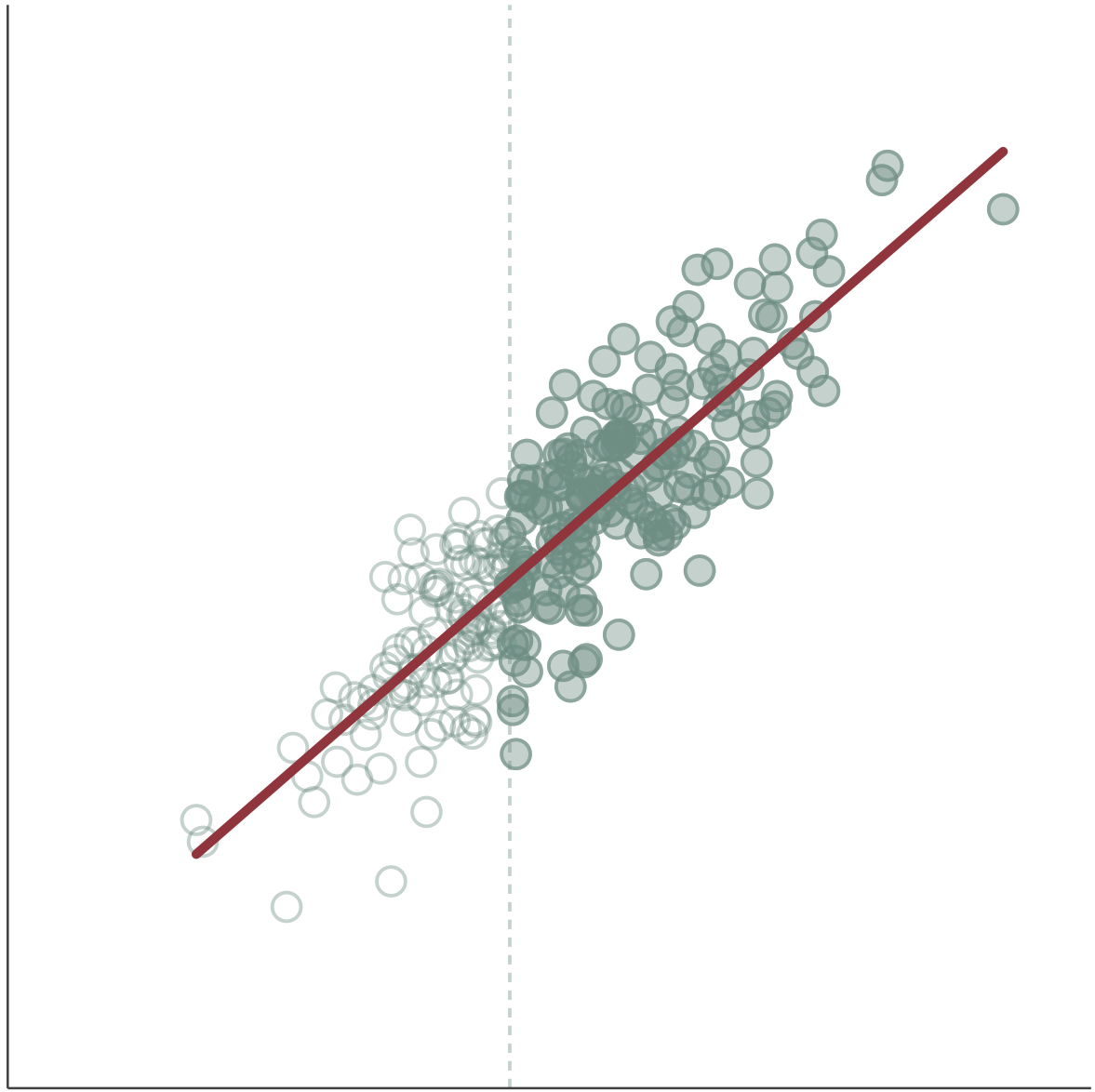
- True population relationship



UNIT NONRESPONSE

Selection on x

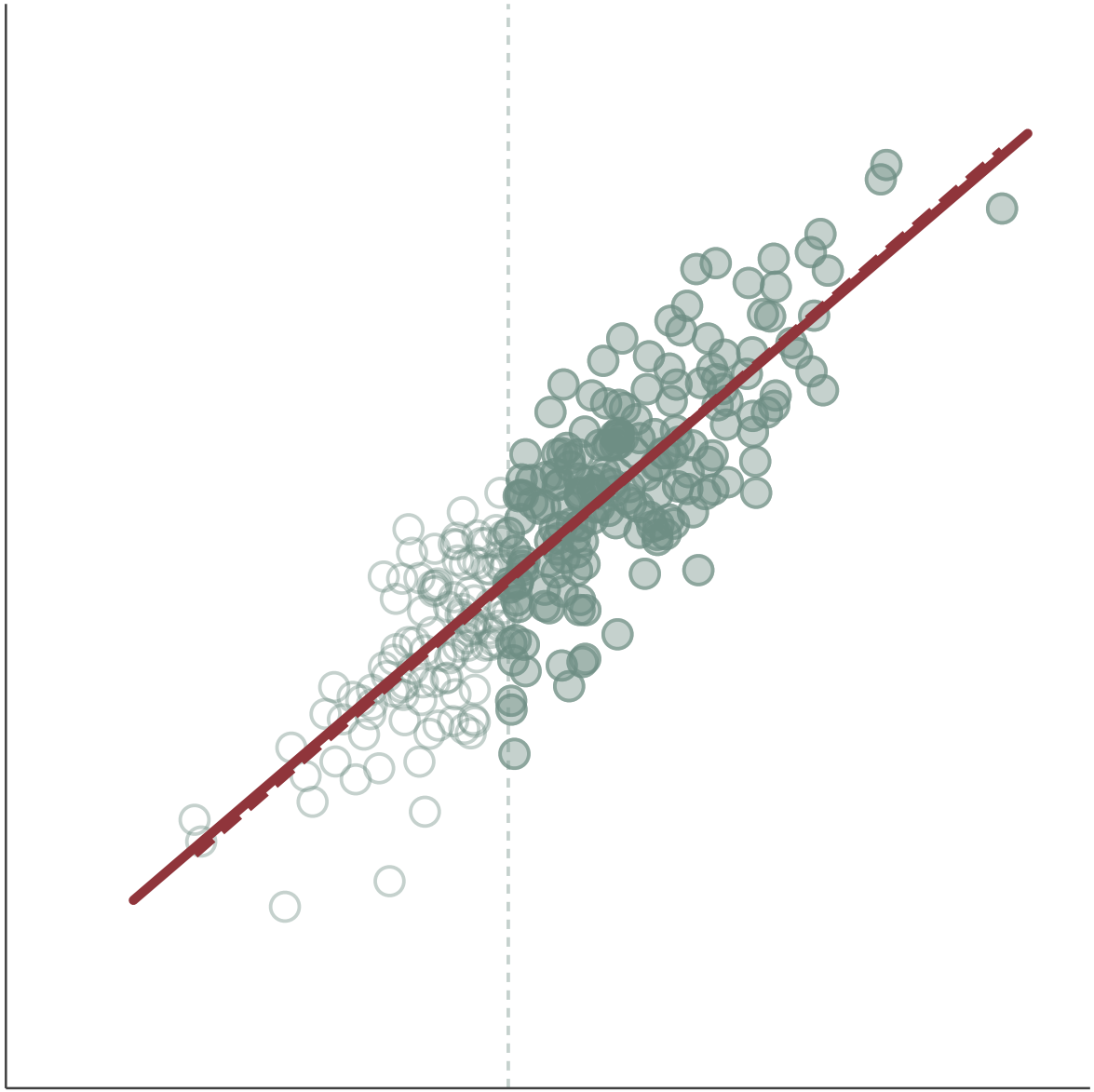
- True population relationship
- Estimated only observing those with high x



UNIT NONRESPONSE

Selection on x

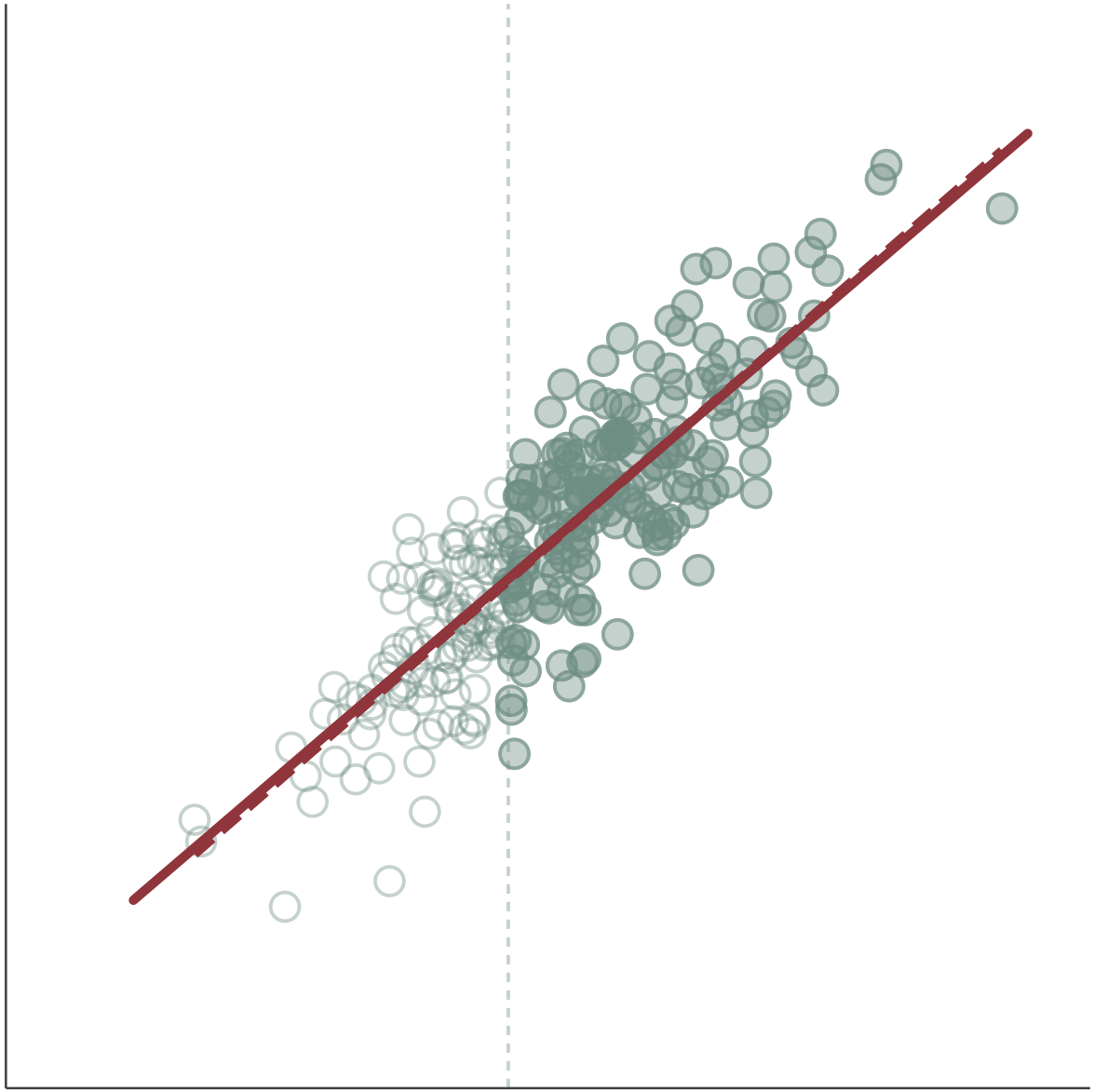
- True population relationship
- Estimated only observing those with high x
- Amazingly, no bias in this case



UNIT NONRESPONSE

Selection on x

- True population relationship
- Estimated only observing those with high x
- Amazingly, no bias in this case
- But: smaller n, larger sampling variance



ITEM NONRESPONSE

Some notation

z = Variable with missing data

X = All other variables

R_z = 1 if z is missing, 0 if observed

"Missing completely at random" (MCAR): Missingness does not depend on z or any other variable

$$\Pr(R_z = 1|X, z) = \Pr(R_z = 1)$$

Can this assumption be tested?

- Yes, look at mean differences in X by R_z or do a regression of X on R_z
- However, not possible to test whether R_z depends on z (would require observing the missing values)

"Missing at random" (MAR): Missingness does not depend on z conditional on other variables

$$\Pr(R_z = 1|X, z) = \Pr(R_z = 1|X)$$

Can this assumption be tested?

- No, again, we don't know whether R_z depends on z without observing the missing values

"Not missing at random" (NMAR): Missingness depends on z, even conditional on other variables

$$\Pr(R_z = 1|X, z) \neq \Pr(R_z = 1|X)$$

Note: no existing method handles NMAR except in very special circumstances

BAD SOLUTIONS

Casewise Deletion

- Default in Stata and other software (”complete case analysis”)
- All observations with missing data are dropped
- Can lead to very large loss if the number of variables is large, cumulative nonresponse
- Also leads to bias unless data are MCAR

Mean dummy imputation

- Replace data on missings with variable means and include an indicator (0/1) for imputed
- Was common in the past but in fact leads to severe biases

Regression imputation

- Replace data on missings with conditional means predicted from other variables
- Better, but underestimates variance and leads to overconfidence

BETTER SOLUTIONS

A good method to deal with missing data should

- Minimise bias in parameter estimates
- Make maximum use of available data
- Give an accurate reflection of uncertainty (SE:s, CI:s, p-values)

Multiple imputation (MI)

- Like regression imputation, but adds a random error to the prediction and generates M different datasets (usually 5)
- The model is estimated M times on imputed data and results averaged

Full information maximum likelihood (FIML)

- Like regression imputation, but incorporates missing values into regression likelihood function and runs just once

Both methods assume that data are MCAR or MAR and they are consistent under the same assumptions

Multiple imputation (MI)

1. Specify how to save imputations (wide or long format)

```
mi set wide
```

2. Specify the variables to be imputed

```
mi register imputed x1 x2
```

3. Specify the variables with no missing (optional)

```
mi register regular y x3 x4 x5
```

4. Select imputation model. The simplest is `mi impute mvn` which assumes that all data are multivariate normal

```
mi impute mvn x1 x2 = y x3 x4 x5, add(20)
```

5. Good, now estimate the model!

```
mi estimate: regress y x1 x2 x3 x4 x5
```

MULTIPLE IMPUTATION (MI)

Multivariate normality is not always appropriate to assume

An alternative is multivariate imputation by chained equations (MICE) which lets you specify the correct functional form for each variable; in Stata, `mi impute chained`

We have missing data on gender (binary), ethnicity (nominal), and age (interval). We will use logit for imputing gender, multinomial for ethnicity, linear regression for age

The initial steps are the same as before

```
mi set wide
mi register imputed female ethnicity age
```

To impute we type

```
mi impute chained (regress) age (logit) female
  (mlogit) ethnicity = y x3 x4 x5, add(20)
```

FULL INFORMATION MAXIMUM LIKELIHOOD (FIML)

In this case there is no need to specify anything in advance, Stata finds the missing values for you

The syntax is the same as for `sem` that we have seen before, just add option `method(mlmv)` (i.e. maximum likelihood with missing values)

To estimate the example from before we would simply have to type:

```
sem (y <- x1 x2 x3 x4 x5), method(mlmv)
```

Instead of

```
mi set wide
mi register imputed x1 x2
mi register regular y x3 x4 x5
mi impute mvn x1 x2 = y x3 x4 x5, add(20)
mi estimate: regress y x1 x2 x3 x4 x5
```

WHICH METHOD?

Advantages with maximum likelihood

- More efficient than multiple imputation
- Always produces same result whereas MI is random
- Imputation and estimation in one model; no conflict
- Fewer decisions about the ”correct” imputation model

Advantages with multiple imputation

- Flexible, can be combined with any model or software
- Can incorporate information outside estimation model
- No need to assume multivariate normality

Estimation $\neq$ prediction

- Imputation is prediction, we tolerate ”kitchen sink” models
- Multicollinearity matters less in prediction than in estimation
- Assumed normality may matter more in prediction than in estimation

SUMMARY

Like measurement error, missing data is also a cause for (mostly) underestimation

Missing data methods are generally more developed, accessible, and widely adopted than corrections for measurement error

But as always there are assumptions, important to treat results with some healthy scepticism

READINGS

Recommended:

- Pepper, John, Carol Petrie, and Sean Sullivan. 2010. "Measurement error in criminal justice data." Handbook of Quantitative Criminology. Springer.
- Allison, Paul D. 1999. "Missing Data" In: SAGE Handbook of Quantitative Methods in Psychology.

Optional:

- Bollen, Kenneth A. 2012. "Instrumental variables in sociology and the social sciences." Annual Review of Sociology 38: 37-72.

Questions?