Practical Data Issues

Department of Political Science and Government
Aarhus University

March 3, 2015

Tidy Data Activity

- Construct the data described on the worksheet into a rectangular dataset
- You can use Stata's data editor, Excel, a Word table, etc.
- You have 10 minutes

Data Formats

- A dataset is a rectangular matrix of:
  - Observations (rows)
  - Variables (columns)
- But what counts as an observation?
  - Countries, by year

Data Formats

- A dataset is a rectangular matrix of:
  - Observations (rows)
  - Variables (columns)
- But what counts as an observation?
  - Countries, by year
  - Dyads (e.g., couples, neighboring countries)

Data Formats

- A dataset is a rectangular matrix of:
  - Observations (rows)
  - Variables (columns)

- But what counts as an observation?
  - Countries, by year
  - Dyads (e.g., couples, neighboring countries)
  - Candidates, by election

Data in Wide Format

Country	GDP 2012	GDP 2013	Life Exp. 2012	Life Exp. 2013
Afghanistan	20.5	20.3	60	61
Albania	12.3	12.9	77	77
Algeria	204.3	210.2	71	71
Angola	114.3	124.2	51	51

Data in Long Format

Country	Year	GDP ($B)	Life Expectancy
Afghanistan	2012	20.5	60
Afghanistan	2013	20.3	61
Albania	2012	12.3	77
Albania	2013	12.9	77
Algeria	2012	204.3	71
Algeria	2013	210.2	71
Angola	2012	114.3	51
Angola	2013	124.2	51

Wide and Long in Stata

- When data are cross-sectional, there is only wide
- When data have other forms, they can be represented in multiple ways
- Next week we'll start discussing over-time data
- In most case, we need data in **long** or "tidy" format
- In Stata, this will require the **reshape** and **xtset** commands

Merging Multiple Datasets

- We can only analyze one dataset at a time
- All data about our observations needs to be in one file
- Often we need data from multiple files together
- To use them, we need to merge
- In Stata, we use the **merge** command

Four Ways of Merging Data

1. 1:1
2. 1:many
3. many:1
4. many:many

1:1 Merging

Country	GDP/cap
Afghanistan	665
Denmark	59,832
USA	53,042

Dataset A

Country	Region
Afghanistan	Asia
Denmark	Europe
USA	Americas

Dataset B

1:many Merging

Person	Country
11	Afghanistan
12	Denmark
13	Denmark
14	USA
14	USA

Dataset A

Country	Region
Afghanistan	Asia
Denmark	Europe
USA	Americas

Dataset B

many:1 Merging

Country	Region
Afghanistan	Asia
Denmark	Europe
USA	Americas

Dataset A

Person	Country
11	Afghanistan
12	Denmark
13	Denmark
14	USA
14	USA

Dataset B

Aggregating Observations

- Our data are not always available on the unit of analysis we need
- For example, we might have multiple observations (rows) for a given unit and we need to create a dataset that just has one observation (row) for each unit
- In Stata, we use the **collapse** command
- Examples?

Aggregating Observations

Person	Country	Age
11	Afghanistan	45
12	Denmark	42
13	Denmark	56
14	USA	31
14	USA	75

Dataset A

Country	Mean Age
Afghanistan	45
Denmark	49
USA	53

Dataset B

Questions about merging or aggregation?

Scale Constructions

- Scale construction is the act of aggregating multiple variables into a smaller number of variables
- Two advantages:
  - Reducing measurement error
  - Avoiding collinearity
- Examples

Scale Constructions

- Scale construction is the act of aggregating multiple variables into a smaller number of variables
- Two advantages:
  - Reducing measurement error
  - Avoiding collinearity
- Examples
  -

Scale Constructions

- Scale construction is the act of aggregating multiple variables into a smaller number of variables
- Two advantages:
  - Reducing measurement error
  - Avoiding collinearity
- Examples
  -
  - Others

Cronbach's α

- Scaling only makes sense if variables "go together"
  - We can assess them *pairwise* by looking at correlations between variables
  - But it's helpful to have a way to assess the scale as a whole
- Definition: $\alpha = \dfrac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$
  - N: number of items
  - $\bar{c}$: average covariance of items
  - $\bar{v}$: average variance of items

Cronbach's α in Stata I

```
. corr price headroom trunk weight length, cov
(obs=45)

             |    price headroom    trunk   weight   length
-------------+---------------------------------------------
       price |  7.7e+06
    headroom |  511.298  .718434
       trunk |  3793.78  2.78662  20.4343
      weight |  1.2e+06  395.896  2544.42   658519
      length |  28383.1  12.5265  79.5833  18332.8  561.391
```

Cronbach's α in Stata II

```
. alpha price headroom trunk weight length, item

Test scale = mean(unstandardized items)

                                                            average
                              item-test     item-rest      interitem
Item         |  Obs  Sign   correlation   correlation     covariance      alpha
-------------+-----------------------------------------------------------------
price        |   70    +       0.9360        0.2201        3120.148      0.0854
headroom     |   66    +       0.2471        0.2453        182861.2      0.2626
trunk        |   69    +       0.3928        0.3752        186996.9      0.2649
weight       |   64    +       0.5665        0.3710        5565.038      0.0106
length       |   69    +       0.5604        0.5414        180695.7      0.2578
-------------+-----------------------------------------------------------------
Test scale   |                                             111565.7      0.2483
-------------------------------------------------------------------------------
```

Other forms of scaling

Other forms of scaling

- IRT Models

Other forms of scaling

- IRT Models
- Factor Analysis

Other forms of scaling

- IRT Models
- Factor Analysis
- Principal Components Analysis

Other forms of scaling

- IRT Models
- Factor Analysis
- Principal Components Analysis

- We are not discussing these here
- You can achieve them in Stata's SEM module

Questions about scaling?

Summary: Data Preparation

- This course is about statistics and data analysis
- Data preparation is often the most time consuming part of research
- Data are rarely, if ever, in the form you need to analyze them properly

Data "Cleaning"

- Once data are structured the way we need them, we're still not done
- Data "cleaning" is the final step before analysis:
  - Recoding variables
  - Typo and coding corrections
  - Scale construction
  - Handling missing data

Missing Data Ideal World

- Everything we've talked about assumes no missingness
- We always assume that we have a representative sample of i.i.d. data
- We analyze all of our data as is

Missing Data

- What is it?

Missing Data

- What is it?
- Why are data missing?

Missing Data

- What is it?
- Why are data missing?
- How often do we encounter missing data?

Does Missingness Matter?

If the data are missing at random, then the size of the random sample available from the population is simply reduced. Although this makes the estimators less precise, it does not introduce any bias [. . .] There are ways to use the information on observations where only some variables are missing, but this is not often done in practice. The improvement in the estimators is usually slight, while the methods are somewhat complicated. In most cases, we just ignore the observations that have missing information. (Wooldridge 2013, 314)

Missing Data in Practice

- Often have missing data for a variety of reasons
- We often don't realize we have missing data
- Missing data can be problematic (but not always)
- Stata handles missingness through **complete case** or **available case** analysis

Complete/Available Cases

- **Complete case analysis** involves subsetting a dataset to retain only observations that are complete on all variables before any analysis
- **Available case analysis** involves dynamically subsetting a dataset to retain only observations that are complete on all variables used in a given analysis
  - Sometimes also called *case-wise deletion* or *list-wise deletion*

Complete/Available Cases

- **Complete case analysis** involves subsetting a dataset to retain only observations that are complete on all variables before any analysis
- **Available case analysis** involves dynamically subsetting a dataset to retain only observations that are complete on all variables used in a given analysis
  - Sometimes also called *case-wise deletion* or *list-wise deletion*
- Do we use either of these techniques?

Impacts of Missingness

1. Scale construction problems
2. Statistical efficiency
3. Representativeness (External validity)
4. Comparability of subsample analyses
5. Causal inference

Possible Impact 1: Scales

- It is common to analyze variables constructed as scales
  - Simple additive scales being the most common
- Examples?
  - Political knowledge
  - Frequency of voting
  - Democracy
  - Budgets across multiple domains

A Simple Example

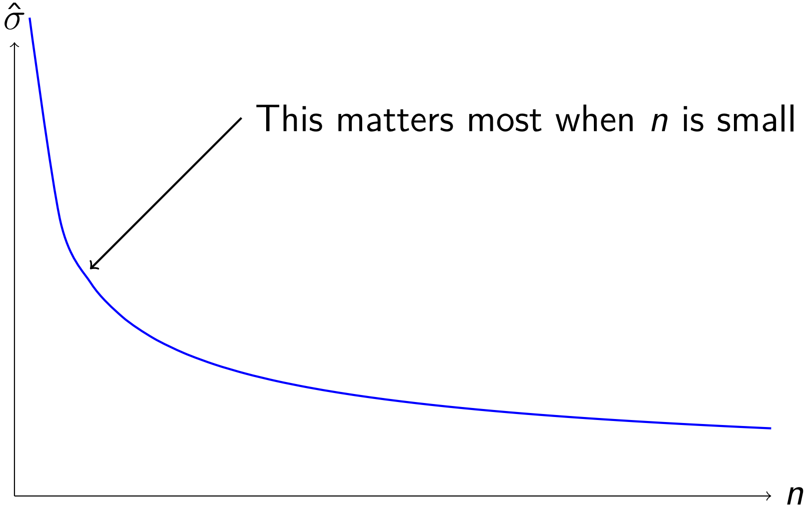
Case	Item 1	Item 2	Item 3	Sum
A	1	2	1	?
B	1	.	3	?
C	.	1	1	?
D	2	1	2	?
E	1	.	.	?
F	.	.	.	?

Possible Impact 1: Scales

- When constructing multi-item scales, we need to know how to deal with missingness
- Stata's default is to coerce missingness to zero
- Another strategy is *imputation*

Possible Impact 2: Efficiency

- Recall: $Var(\hat{\beta}) = \hat{\sigma}(\mathbf{X}'\mathbf{X})^{-1}$
- And $\hat{\sigma}^2 = \frac{SSR}{n-2}$, so that $\hat{\sigma} = \frac{\sqrt{SSR}}{\sqrt{n-2}}$
- As sample size increases we gain precision
- Missing data reduces our *effective sample size* for analysis


$\hat{\sigma}$
This matters most when n is small
n

Possible Impact 3: Representativeness

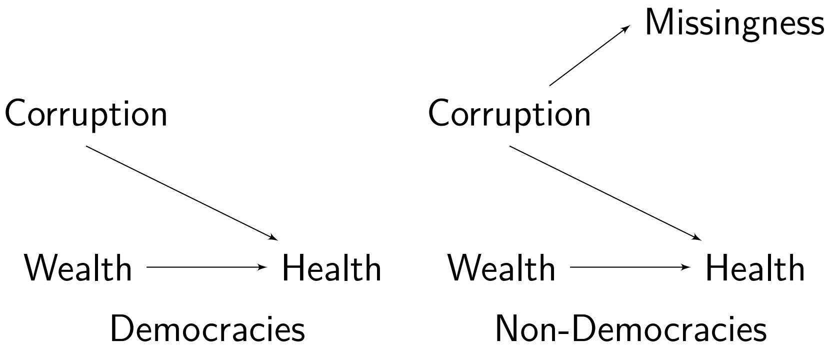
- Recall: We generally try to make inferences from sample to a well-specified population
- If missingness is *ignorable*, we simply have a smaller sample
- If missingness is not *ignorable*, we no longer have a representative sample
  - This leads to bias in our estimates

Possible Impact 4: Comparability

- When there is missingness, we (and Stata) default to *available case analysis*
- Our analyses might be based on different subsamples of our data
- Thus the precision of our estimates from different analyses might vary
- Can be solved through *complete case analysis*

Possible Impact 5: Causal Inference

- Our inferences might be biased if missingness is caused by a third variable
- This is especially bad if the third variable is also causally important for our outcome


Missingness
Corruption
Wealth
Health
Democracies
Corruption
Wealth
Health
Non-Democracies

Impact of Missingness Depends on *Why* Data Are Missing

- Missing Completely At Random (MCAR)
- Missing At Random
- Missing Not At Random (MNAR)

MCAR/Ignorable

- Best-case scenario
- Our data constitute a representative subsample of our sample, making it a representative sample of our population
- Examples?

MCAR/Ignorable

- Best-case scenario
- Our data constitute a representative subsample of our sample, making it a representative sample of our population
- Examples?
  - We obtain a complete sample but randomly analyze only part of it

MCAR/Ignorable

- Best-case scenario
- Our data constitute a representative subsample of our sample, making it a representative sample of our population
- Examples?
    - We obtain a complete sample but randomly analyze only part of it
    - Survey respondents randomly assigned to different questionnaires

MCAR/Ignorable

- Best-case scenario
- Our data constitute a representative subsample of our sample, making it a representative sample of our population
- Examples?
  - We obtain a complete sample but randomly analyze only part of it
  - Survey respondents randomly assigned to different questionnaires
- How do we deal with missingness?

MCAR/Ignorable

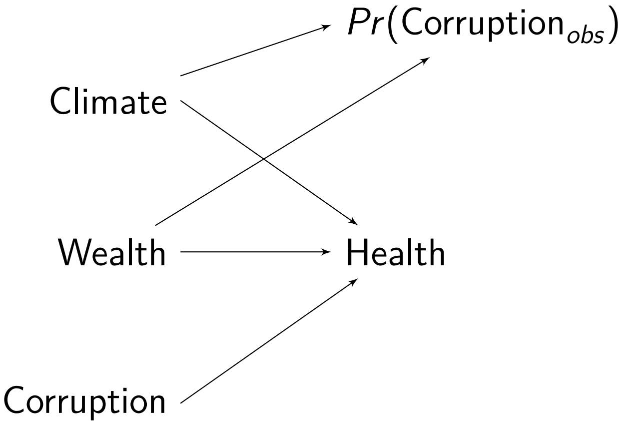
- Best-case scenario
- Our data constitute a representative subsample of our sample, making it a representative sample of our population
- Examples?
  - We obtain a complete sample but randomly analyze only part of it
  - Survey respondents randomly assigned to different questionnaires
- How do we deal with missingness?
  - We can probably ignored it

Impacts of Missingness (MCAR)

1. **Scale construction problems**
2. **Statistical efficiency**
3. Representativeness (External validity)
4. Comparability of subsample analyses
5. Causal inference

MAR

- Middle-ground scenario
- Data are missing for a (non-random) reason that we understand and observe
- Missingness is ***conditionally*** **ignorable**


$Pr(\text{Corruption}_{obs})$
Climate
Wealth
Health
Corruption

Impacts of Missingness (MAR)

1. **Scale construction problems**
2. **Statistical efficiency**
3. Representativeness (External validity)
4. **Comparability of subsample analyses**
5. Causal inference

Handling MAR Data

- Regression adjustment
- Reweighting
- Single imputation
    - Several possible methods
- Multiple imputation
    - Several possible methods

Regression Adjustment

- If missingness only depends on right-hand side variables in our model, then regression alone with adjust for missingness and yield *unbiased* coefficient estimates
- We still lose efficiency because of the missing observations

Regression Adjustment

- If missingness only depends on right-hand side variables in our model, then regression alone with adjust for missingness and yield *unbiased* coefficient estimates
- We still lose efficiency because of the missing observations
- Caution: A sometimes-common practice
  - Include an indicator variable for missingness in X
  - Regress Y on X and the X_obs indicator
  - Tends to produce biased estimates

Weighting adjustments

- Stratify the sample based on observed characteristics, where the proportion of the *population* in each stratum is also known
- Reweight each observation so sample matches population distributions
  - Essentially, over-weight observed cases from strata where there are missing values
- Several variants of this:
  - Weighting classes
  - Post-stratification
  - Raking

Single imputation

- Fill in missing values with an *imputed* value
- Several different methods, including:
  - Zero
  - Mean value
  - Random value
  - Inferred value
  - Hot-Deck imputation
  - Regression imputation

Single Imputation I

- Zero: Will bias results, unless $\bar{X} = 0$

Single Imputation I

- Zero: Will bias results, unless $\bar{X} = 0$
- Mean: Unbiased...why?

Single Imputation I

- Zero: Will bias results, unless $\bar{X} = 0$
- Mean: Unbiased...why?
- Random: Unbiased...why?

Single Imputation I

- Zero: Will bias results, unless $\bar{X} = 0$
- Mean: Unbiased...why?
- Random: Unbiased...why?
- Inferred
  - Uses observed data to guess at missing value
  - Could be historical records, logic, etc.

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
- Regression Imputation

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
  2. For every missing value, carry forward last observed value

- Regression Imputation

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
  2. For every missing value, carry forward last observed value
  3. Imputations depend on sort order
- Regression Imputation

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
  2. For every missing value, carry forward last observed value
  3. Imputations depend on sort order

- Regression Imputation
  - Regress partially observed variable on all complete variables

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
  2. For every missing value, carry forward last observed value
  3. Imputations depend on sort order

- Regression Imputation
  - Regress partially observed variable on all complete variables
  - Replace missing value with fitted value $\hat{y}$ from regression

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
  2. For every missing value, carry forward last observed value
  3. Imputations depend on sort order

- Regression Imputation
  - Regress partially observed variable on all complete variables
  - Replace missing value with fitted value $\hat{y}$ from regression
  - Imputations depend on model

Single Imputation I

- Hot-Deck Imputation
  1. Sort dataset by all complete variables
  2. For every missing value, carry forward last observed value
  3. Imputations depend on sort order

- Regression Imputation
  - Regress partially observed variable on all complete variables
  - Replace missing value with fitted value $\hat{y}$ from regression
  - Imputations depend on model
  - Can dramatically overstate certainty unless a stochastic component is added

Multiple Imputation

- Apply a stochastic single imputation technique multiple times and merge the results of the analysis performed on each imputed dataset
  - Usually some form of regression imputation
- Attempts to account for uncertainty due to imputation
  - Single imputation overstates our certainty

MI Procedure

1. Impute missing values and estimate $\hat{\beta}_m$
2. Repeat for all M datasets
3. Aggregate results: $\hat{\beta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m$
4. Account for missingness when estimating variance:
   - Within $= \frac{1}{m}\sum_1^M Var(\hat{\beta}_m)$
   - Between $= \frac{1}{m-1}\sum_1^M(\hat{\beta}_m - \hat{\beta})^2$
   - $Var(\hat{\beta}) = \text{Within} + (1 + \frac{1}{m})\text{Between}$

An Example I

- What is the effect of university education on an individuals' political tolerance?
- Missingness in various covariates
- Multiply impute missing values
- On each imputed dataset, we estimate: Tolerance $= \beta_0 + \beta_1$Education $+ \beta_{2 \ldots k}$Controls
- Our test statistic is $\hat{\beta}_{\text{Education}}$

An Example II

Dataset	$\hat{\beta}_{\text{Education}}$	$SE_{\hat{\beta}}$	$Var(\hat{\beta})$
1	4.32	0.95	0.9025
2	4.15	1.16	1.3456
3	4.86	0.83	0.6889
4	3.98	1.04	1.0816
5	4.50	0.91	0.8281

An Example II

Dataset	$\hat{\beta}_{\text{Education}}$	$SE_{\hat{\beta}}$	$Var(\hat{\beta})$
1	4.32	0.95	0.9025
2	4.15	1.16	1.3456
3	4.86	0.83	0.6889
4	3.98	1.04	1.0816
5	4.50	0.91	0.8281

$$\hat{\beta}_{Overall} = \frac{4.32+4.15+4.86+3.98+4.50}{5} = 4.362$$

An Example III

$$Var_{\mathrm{W}} = \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281)$$

An Example III

$$\begin{aligned} Var_{\mathrm{W}} &= \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281) \\ &= \frac{4.8467}{5} = 0.96934 \end{aligned}$$

An Example III

$$
\begin{aligned}
Var_{\mathrm{W}} &= \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281) \\
&= \frac{4.8467}{5} = 0.96934
\end{aligned}
$$

$$
Var_{\mathrm{B}} = \frac{1}{m-1}(-0.042^2 + 0.212^2 + 0.498^2 + -0.382^2 + 0.138^2)
$$

An Example III

$$
\begin{aligned}
Var_{\mathrm{W}} &= \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281) \\
&= \frac{4.8467}{5} = 0.96934
\end{aligned}
$$

$$
\begin{aligned}
Var_{\mathrm{B}} &= \frac{1}{m-1}(-0.042^2 + 0.212^2 + 0.498^2 + -0.382^2 + 0.138^2) \\
&= \frac{0.45968}{4} = 0.11492
\end{aligned}
$$

An Example III

$$
\begin{aligned}
Var_{\mathrm{W}} &= \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281) \\
&= \frac{4.8467}{5} = 0.96934
\end{aligned}
$$

$$
\begin{aligned}
Var_{\mathrm{B}} &= \frac{1}{m-1}(-0.042^2 + 0.212^2 + 0.498^2 + -0.382^2 + 0.138^2) \\
&= \frac{0.45968}{4} = 0.11492
\end{aligned}
$$

$$
Var(\hat{\beta}) = W + (1 + \frac{1}{m})B = 1.107244
$$

An Example III

$$Var_{\mathrm{W}} = \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281)$$
$$= \frac{4.8467}{5} = 0.96934$$

$$Var_{\mathrm{B}} = \frac{1}{m-1}(-0.042^2 + 0.212^2 + 0.498^2 + -0.382^2 + 0.138^2)$$
$$= \frac{0.45968}{4} = 0.11492$$

$$Var(\hat{\beta}) = W + (1 + \frac{1}{m})B = 1.107244$$

An Example III

$$
\begin{aligned}
Var_{\mathrm{W}} &= \frac{1}{5}(0.9025 + 1.3456 + 0.6889 + 1.0816 + 0.8281) \\
&= \frac{4.8467}{5} = 0.96934
\end{aligned}
$$

$$
\begin{aligned}
Var_{\mathrm{B}} &= \frac{1}{m-1}(-0.042^2 + 0.212^2 + 0.498^2 + -0.382^2 + 0.138^2) \\
&= \frac{0.45968}{4} = 0.11492
\end{aligned}
$$

$$
Var(\hat{\beta}) = W + (1 + \frac{1}{m})B = 1.107244
$$

$$
SE(\hat{\beta}) = \sqrt{1.107244} = 1.052257
$$

MAR: Conclusion

- If we assume MAR, lots of available strategies

MAR: Conclusion

- If we assume MAR, lots of available strategies
- Added value of imputation depends on scale of missingness and assumptions

MAR: Conclusion

- If we assume MAR, lots of available strategies
- Added value of imputation depends on scale of missingness and assumptions
- Can overstate our certainty about model estimates

MAR: Conclusion

- If we assume MAR, lots of available strategies
- Added value of imputation depends on scale of missingness and assumptions
- Can overstate our certainty about model estimates
- Can introduce measurement error if we misunderstand the pattern of missingness, which then leads to bias

Questions about MAR?

MNAR

- Worst-case scenario
- Missingness is **non-ignorable**
- Data are missing due to factors that are in our model
- Examples?

MNAR

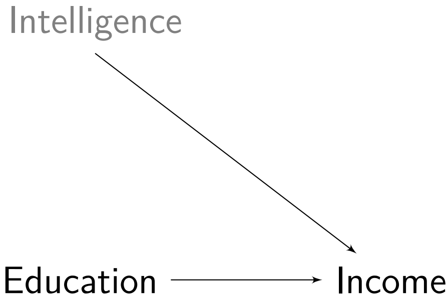
- Worst-case scenario
- Missingness is **non-ignorable**
- Data are missing due to factors that are in our model
- Examples?
  - Survey participation based on topic

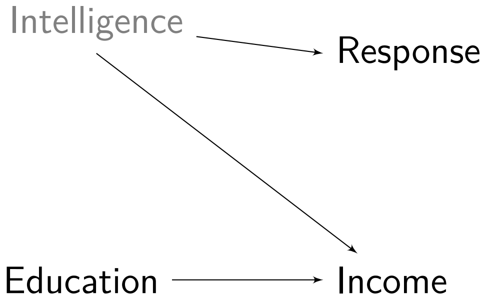
MNAR

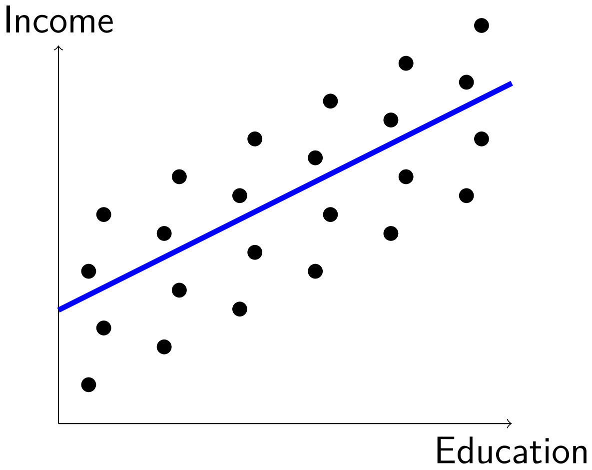
- Worst-case scenario
- Missingness is **non-ignorable**
- Data are missing due to factors that are in our model
- Examples?
  - Survey participation based on topic
  - Income reporting based on income

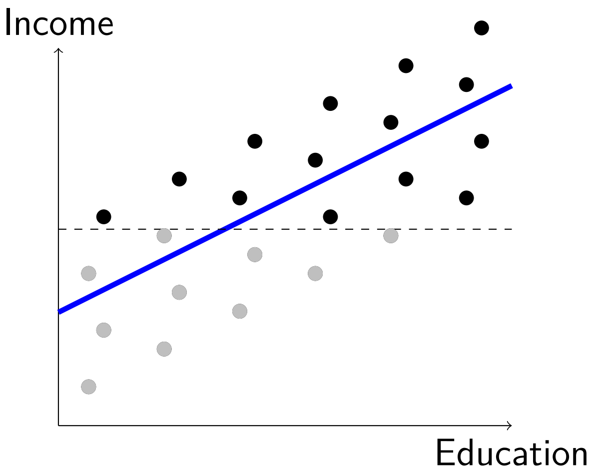
MNAR: Special Cases

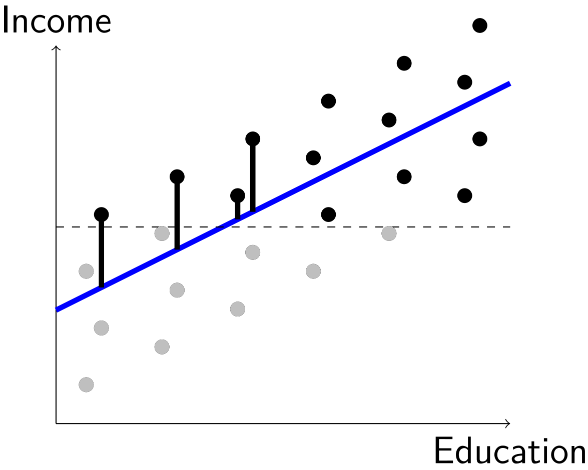
1. Truncation (Sample selection bias)
2. Censoring

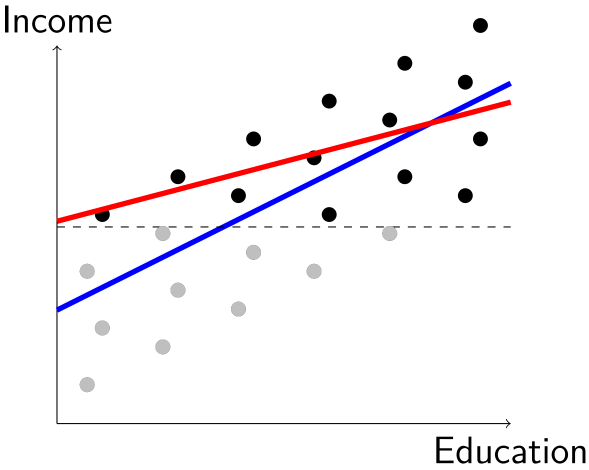

Intelligence
Education
Income


Intelligence
Response
Education
Income


Income
Education


Income
Education


Income
Education


Income
Education

Sample Selection Bias

- Whether we observe a unit depends on factors related to X and Y
- Common analytic strategy: Heckman Models
  - OLS with an additional covariate
  - Regress missingness on variable(s) *not* in our main model
  - Include predicted probability of observing case as a covariate in main model

Sample Selection Bias

- Whether we observe a unit depends on factors related to X and Y
- Common analytic strategy: Heckman Models
  - OLS with an additional covariate
  - Regress missingness on variable(s) *not* in our main model
  - Include predicted probability of observing case as a covariate in main model
- Examples?

Sample Selection Bias

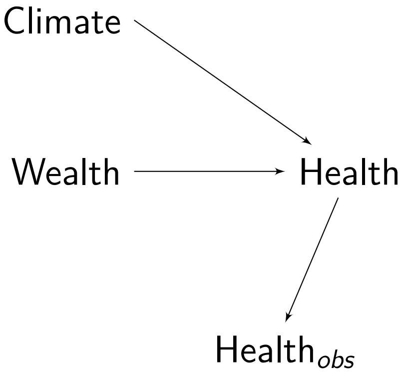
- Whether we observe a unit depends on factors related to X and Y
- Common analytic strategy: Heckman Models
  - OLS with an additional covariate
  - Regress missingness on variable(s) *not* in our main model
  - Include predicted probability of observing case as a covariate in main model
- Examples?
  - Effect of grades on job performance

Sample Selection Bias

- Whether we observe a unit depends on factors related to X and Y
- Common analytic strategy: Heckman Models
  - OLS with an additional covariate
  - Regress missingness on variable(s) *not* in our main model
  - Include predicted probability of observing case as a covariate in main model
- Examples?
  - Effect of grades on job performance
  - Systematic survey nonresponse

MNAR: Special Cases

1. Truncation (Sample selection bias)
2. Censoring


Climate
Wealth
Health
Health$_{obs}$

Censoring

- Values of Y above (or below) a threshold are scored at the threshold value
- Sometimes “top-coding” or “bottom-coding”
- Basically systematic measurement error
- Examples?

Censoring

- Values of Y above (or below) a threshold are scored at the threshold value
- Sometimes "top-coding" or "bottom-coding"
- Basically systematic measurement error
- Examples?
  - Income self-reports on surveys

Censoring

- Values of Y above (or below) a threshold are scored at the threshold value
- Sometimes “top-coding” or “bottom-coding”
- Basically systematic measurement error
- Examples?
  - Income self-reports on surveys
  - Measurement tool insensitive below threshold

Dealing with censoring

- Estimate a modified regression model
- OLS cannot accommodate this
- Common approach is the **Tobit model**
- We'll talk about this later

Impacts of Missingness (MNAR)

1. **Scale construction problems**
2. **Statistical efficiency**
3. **Representativeness (External validity)**
4. **Comparability of subsample analyses**
5. **Causal inference**

The Big Problem

- Choosing among MCAR vs. MAR vs. MNAR is an untestable assumption
- We usually never *completely* know why data are missing

Questions about missing data?

Activity Answer: Wide Format

CNTRY	UNE10	FEM10	MIL14	UNEM14	FEM14	DEM14	HDI14	EDU14	HIV14	IMR14	LE14
AFGH	.	28	4.37	.	28	1	0.453	9	0.1	74.5	48.28
COL	12	13	1.99	11.6	12	1	.	.	.	.	.
CUB	.	43	.	.	49	0	0.824	9	0.2	4.6	78.96
DEN	7.5	38	0.64	7.6	39	1	0.898	10	0.2	3.3	79.09
EGY	.	.	.	.	.	0	.	.	.	.	.
NK	4.5	16	9.31	4.4	16	0	.	.	.	.	.
UK	7.9	22	.	7.8	23	1	.	.	.	.	.
USA	9.7	17	.	.	18	1	.	.	.	.	.

CNTRY	MIL10	UNE10	FEM10	DEM10	HDI10	EDU10	HIV10	IMR10	LE10	MIL14	UNEM14	FEM14	DEM14	HDI14	EDU14	HIV14	IMR14	LE14
AFGH	.	.	28	1	.	.	.	.	.	4.37	.	28	1	0.453	9	0.1	74.5	48.28
COL	.	12	13	1	.	.	.	.	.	1.99	11.6	12	1	.	.	.	.	.
CUB	.	.	43	0	.	.	.	.	.	.	.	49	0	0.824	9	0.2	4.6	78.96
DEN	.	7.5	38	1	.	.	.	.	.	0.64	7.6	39	1	0.898	10	0.2	3.3	79.09
EGY	.	.	.	0	.	.	.	.	.	.	.	.	0	.	.	.	.	.
NK	.	4.5	16	0	.	.	.	.	.	9.31	4.4	16	0	.	.	.	.	.
UK	.	7.9	22	1	.	.	.	.	.	.	7.8	23	1	.	.	.	.	.
USA	.	9.7	17	1	.	.	.	.	.	.	.	18	1	.	.	.	.	.

Activity Answer: Long Format

COUNTRY	YEAR	MIL	UNEMPLO	FEMALE	DEM	HDI	EDU	HIV	IMR	LE
Afghanistan	2014	4.37	.	28	1	0.453	9	0.1	74.5	48.28
Afghanistan	2010	.	.	28	1	.	.	.	.	.
Colombia	2014	1.99	11.6	12	1	.	.	.	.	.
Colombia	2010	.	12	13	1	.	.	.	.	.
Cuba	2014	.	.	49	0	0.824	9	0.2	4.6	78.96
Cuba	2010	.	.	43	0	.	.	.	.	.
Denmark	2014	0.64	7.6	39	1	0.898	10	0.2	3.3	79.09
Denmark	2010	.	7.5	38	1	.	.	.	.	.
Egypt	2014	.	.	.	0	.	.	.	.	.
Egypt	2010	.	.	.	0	.	.	.	.	.
NorthKorea	2014	9.31	4.4	16	0	.	.	.	.	.
NorthKorea	2010	.	4.5	16	0	.	.	.	.	.
UK	2014	.	7.8	23	1	.	.	.	.	.
UK	2010	.	7.9	22	1	.	.	.	.	.
USA	2014	.	.	18	1	.	.	.	.	.
USA	2010	.	9.7	17	1	.	.	.	.	.