Quantitative Social Research II

Workshop 5: Data Quality

Jose Pina-Sánchez

Workshop Aims

- Review the implications of missing data
  - Including wider problems of selection bias
  - And measurement error
- Introduce methods to adjust for missing data
  - Probability weights
  - Imputation

Workshop Aims: Recap

- The assumptions of the linear regression model:
  - Normality: $N \sim (0, Var(e))$
  - Homoskedasticity: $Var(e_i) = Var(e)$
  - Independence: $Cov(e_i, e_j) = 0$
  - No endogeneity: $Cov(X_i, e_i) = 0$
  - **Perfectly measured variables**
  - **No missing data (other than missing at random)**
  - No omitted relevant variables
  - No multicollinearity
  - Linearity

Selection Bias

- Non-probability sampling methods
  - Not every subject in the population has an equal chance of being captured in the sample
  - Tend to produce biased samples
  - I.e. systematically different from the population

Selection Bias

- Non-probability sampling methods
  - Not every subject in the population has an equal chance of being captured in the sample
  - Tend to produce biased samples
  - I.e. systematically different from the population
- Probability sampling methods
  - Everyone has an equal chance, in principle
  - Question: Could probability samples ever be biased?

Selection Bias

- Non-probability sampling methods
  - Not every subject in the population has an equal chance of being captured in the sample
  - Tend to produce biased samples
  - I.e. systematically different from the population
- Probability sampling methods
  - Everyone has an equal chance, in principle
  - Question: Could probability samples ever be biased?
  - Coverage error
  - Non-response

Non-response

- One form of missing data, not the only one
- The most common form of missing data in survey research
- Can take two main forms
  - Unit non-response
  - Item non-response

Non-response

- One form of missing data, not the only one
- The most common form of missing data in survey research
- Can take two main forms
  - Unit non-response
  - Item non-response
- Missing data mechanisms can be classified in three groups
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - Missing not at random (MNAR)
- Different implications depending on the <u>ignorability</u> of the missing data mechanism

Unit and Item Non-Response

ID	Offence type	Seriousness	Prev. convictions	Sentence length
1	ABH		7	18
2	ABH	3	1	5
3	Affray	2	9	12
4				
5	Affray		15	6
6	GBH	1	0	24

Missing Data Mechanisms

- Missing completely at random
  - The missing data mechanism is not related to any of our variables
  - E.g. some of the data was lost by accident
  - Implications: loss of statistical power because of a smaller sample

Missing Data Mechanisms

- Missing completely at random
  - The missing data mechanism is not related to any of our variables
  - E.g. some of the data was lost by accident
  - Implications: loss of statistical power because of a smaller sample
- Missing data at random
  - Related to one or more of the explanatory variables
  - E.g. male judges might forget to submit their survey forms more commonly than female judges

    Private sector workers might decline to participate in the survey more commonly than public sector workers
  - If left unadjusted will bias our estimates, if adjusted becomes 'ignorable'

Missing Data Mechanisms

- Missing completely at random
  - The missing data mechanism is not related to any of our variables
  - E.g. some of the data was lost by accident
  - Implications: loss of statistical power because of a smaller sample
- Missing data at random
  - Related to one or more of the explanatory variables
  - E.g. male judges might forget to submit their survey forms more commonly than female judges

    Private sector workers might decline to participate in the survey more commonly than public sector workers
  - If left unadjusted will bias our estimates, if adjusted becomes 'ignorable'
- Missing data not at random
  - Related to the outcome variable
  - E.g. harsher judges might try to avoid submitting their forms

    Millionaires are more likely not to report their earnings
  - Cannot be adjusted easily, will bias our estimates

Probability Weights

- Using *weights* we can reflect the scarcity/abundance of certain cases in our sample and obtain a more representative sample
  - We can potentially adjust for multiple problems of selection bias beyond non-response

Probability Weights

- Using *weights* we can reflect the scarcity/abundance of certain cases in our sample and obtain a more representative sample
  - We can potentially adjust for multiple problems of selection bias beyond non-response
- One method to create weights is post-stratification
  - If we know the distribution for one or a set of variables in both the target population and our sample
  - We can calculate weights as a ratio of ratios

Gender	Population Proportion	Sample Proportion	Population/ Sample	Weight
Female	.5	.6	.5 /.6	**.8333**
Male	.5	.4	.5 /.4	**1.25**
Total	1	1		

Probability Weights

- Using *weights* we can reflect the scarcity/abundance of certain cases in our sample and obtain a more representative sample
  - We can potentially adjust for multiple problems of selection bias beyond non-response
- One method to create weights is post-stratification
  - If we know the distribution for one or a set of variables in both the target population and our sample
  - We can calculate weights as a ratio of ratios

Gender	Population Proportion	Sample Proportion	Population/ Sample	Weight
Female	.5	.6	.5 /.6	**.8333**
Male	.5	.4	.5 /.4	**1.25**
Total	1	1		

  - Poststratification weights use a different range (in principle 0 to ∞, although in practice we often use 0.3 to 3) from those used in local regression (0 to 1)
  - A weight of 1 means that the influence of that case in our analyses remains unchanged
  - A weight of 2 means that the case counts as two normal cases (its influence is doubled)
  - A weight of 0.5 means that the case influence is halved

Limitations of Weights

- They are not statistically efficient (increase standard errors)
  - Cases with $W < 1$ are not contributing that much
  - Those with $W > 1$ are contributing more than the typical case without increasing the heterogeneity of the sample
  - Trade-off between accuracy and precision
- Can adjust for multiple variables
  - By combining their categories

    e.g. male-white, male-nonwhite, female-white, female-nonwhite
  - Soon runs out of cases for some categories
- Not so flexible to deal with item non-response
  - Imputation methods are normally used instead

Making the most of the Data

- Cases affected by unit-missing are dropped (case 4 below)

ID	Offence type	Seriousness	Prev. convictions	Sentence length
1	ABH		7	18
2	ABH	3	1	5
3	Affray	2	9	12
4				
5	Affray		15	6
6	GBH	1	0	24

Making the most of the Data

- Cases affected by unit-missing are dropped (case 4 below)

ID	Offence type	Seriousness	Prev. convictions	Sentence length
1	ABH		7	18
2	ABH	3	1	5
3	Affray	2	9	12
4				
5	Affray		15	6
6	GBH	1	0	24

- When analysing a variable affected by item non-response we will have to drop cases with any missing values even if they only affect one of the variables, *listwise deletion* (we will have to drop cases 1 and 5 too)

ID	Offence type	Seriousness	Prev. convictions	Sentence length
2	ABH	3	1	5
3	Affray	2	9	12
6	GBH	1	0	24

Making the most of the Data

- Cases affected by unit-missing are dropped (case 4 below)

ID	Offence type	Seriousness	Prev. convictions	Sentence length
1	ABH		7	18
2	ABH	3	1	5
3	Affray	2	9	12
4				
5	Affray		15	6
6	GBH	1	0	24

- When analysing a variable affected by item non-response we will have to drop cases with any missing values even if they only affect one of the variables, *listwise deletion* (we will have to drop cases 1 and 5 too)

ID	Offence type	Seriousness	Prev. convictions	Sentence length
2	ABH	3	1	5
3	Affray	2	9	12
6	GBH	1	0	24

- Using imputation methods we will be able to use cases affected by item non-response

ID	Offence type	Seriousness	Prev. convictions	Sentence length
1	ABH	2	7	18
2	ABH	3	1	5
3	Affray	2	9	12
5	Affray	1	15	6
6	GBH	1	0	24

Single Imputation

- The simplest methods are based on 'single imputation'
  - Aim to replace each missing data point with a plausible value

Single Imputation

- The simplest methods are based on 'single imputation'
  - Aim to replace each missing data point with a plausible value
- Mean imputation
  - Each missing case replaced by the mean of the observed cases in the same item/variable
  - Allows us to make use of all cases
  - Artificially reduces the standard deviation of the variable imputed and the standard errors of any model where it is used

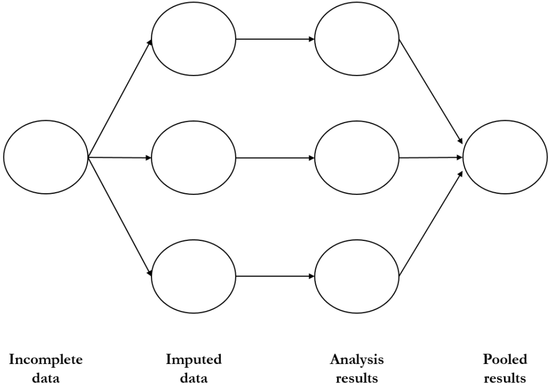
Single Imputation

- The simplest methods are based on 'single imputation'
  - Aim to replace each missing data point with a plausible value
- Mean imputation
  - Each missing case replaced by the mean of the observed cases in the same item/variable
  - Allows us to make use of all cases
  - Artificially reduces the standard deviation of the variable imputed and the standard errors of any model where it is used
- Hot-deck imputation & regression imputation
  - Each missing case replaced with a value from a similar observation in the dataset
  - Uses other variables and cases for which there is complete information to make predictions about the missing values
  - Hot-deck imputation if the prediction is made using matching, regression imputation if using regression
  - Allows using all cases and the effect on the standard deviation will be milder
  - Standard errors still biased from taking the imputed values as data points rather than as estimates for which we are uncertain

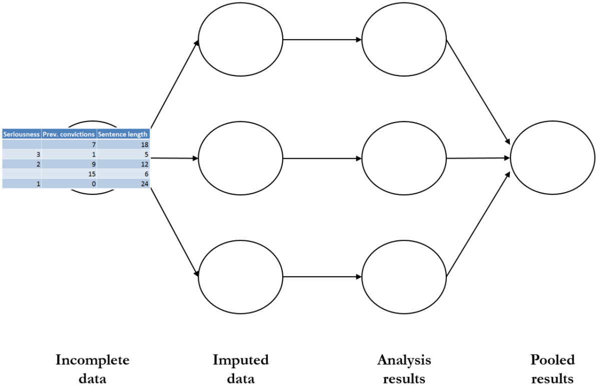
Multiple Imputation

- Multiple imputation
  - Each missing value is replaced with multiple plausible values to generate multiple complete data sets
  - Imputations can be done using regression imputation, hot-deck imputation or similar
  - The analysis is conducted in each of those datasets, results from each analysis are saved and pooled into an average of estimates
  - Having multiple values eliminates the problem of treating imputed cases as real data, i.e. accounts for the uncertainty of the imputation process
  - Generally 3 to 5 imputations are sufficient
  - Computationally intensive

Multiple Imputation

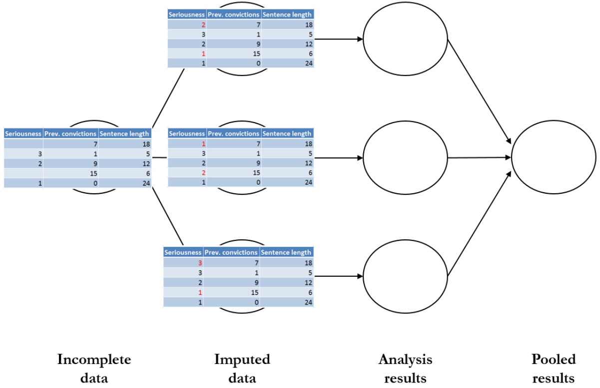


Multiple Imputation

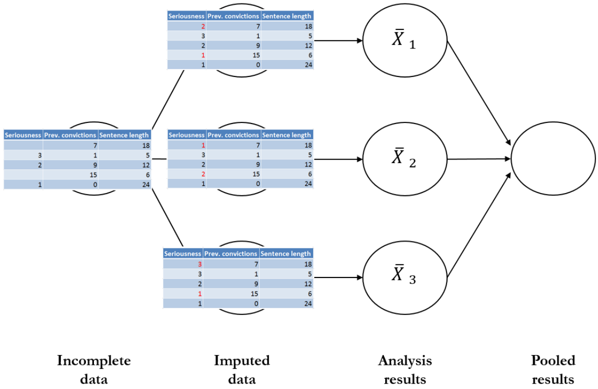


Seriousness	Prev. convictions	Sentence length
	7	18
3	1	5
2	9	12
	15	6
1	0	24

Multiple Imputation



Multiple Imputation



Incomplete data:

Seriousness	Prev. convictions	Sentence length
	7	18
3	1	5
2	9	12
	15	6
1	0	24

Imputed data 1:

Seriousness	Prev. convictions	Sentence length
2	7	18
3	1	5
2	9	12
1	15	6
1	0	24

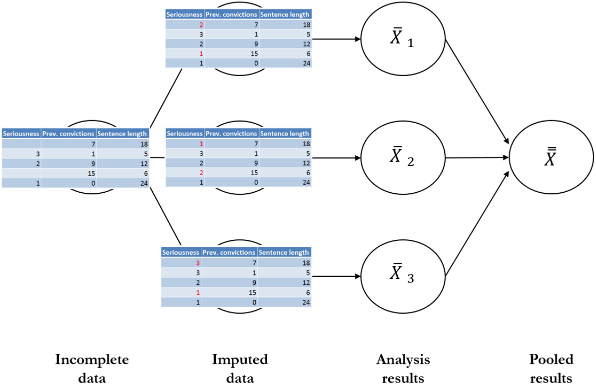
Imputed data 2:

Seriousness	Prev. convictions	Sentence length
1	7	18
3	1	5
2	9	12
2	15	6
1	0	24

Imputed data 3:

Seriousness	Prev. convictions	Sentence length
3	7	18
3	1	5
2	9	12
1	15	6
1	0	24

Multiple Imputation



Seriousness	Prev. convictions	Sentence length
	7	18
3	1	5
2	9	12
	15	6
1	0	24

Seriousness	Prev. convictions	Sentence length
2	7	18
3	1	5
2	9	12
1	15	6
1	0	24

Seriousness	Prev. convictions	Sentence length
1	7	18
3	1	5
2	9	12
2	15	6
1	0	24

Seriousness	Prev. convictions	Sentence length
3	7	18
3	1	5
2	9	12
1	15	6
1	0	24

Measurement Error

- An even more common problem than missing data but hardly ever acknowledged
- Occurs when the *true values* of a variable cannot be obtained
  - $\overbrace{X^*}^{observed} = \overbrace{X}^{true\ value} + \overbrace{\epsilon}^{noise}$
  - can take the form of systematic errors $E(\epsilon) \neq 0$
  - and random errors $E(\epsilon) = 0$

Measurement Error

- An even more common problem than missing data but hardly ever acknowledged
- Occurs when the *true values* of a variable cannot be obtained
  - $\overbrace{X^*}^{observed} = \overbrace{X}^{true\ value} + \overbrace{\epsilon}^{noise}$
  - can take the form of systematic errors $E(\epsilon) \neq 0$
  - and random errors $E(\epsilon) = 0$
- Ubiquitous in all types of quantitative research but specially prevalent in the Social Sciences
  - Survey data affected by memory failures, social desirability (e.g. underreported unemployment, see Pina-Sánchez et al. 2014), etc.
  - Poor operationalisation of concepts (e.g. using earnings to measure poverty; political decentralisation as spending capacity by regional and local governments, see Pina-Sánchez 2014)
  - Measures being played (e.g. arrest goals can inflate crime counts in police data, student satisfaction will increase if I bring chocolates before the module evaluation)
  - Inconsistent raters (e.g. 'blackness' is defined differently by different people, see King & Johnson 2016)

Implications of Measurement Error

- Measurement error adjustments are tricky
  - Either require a 'gold standard' (a subset of our sample for which X is observed)
  - Or to rely on additional assumptions

Implications of Measurement Error

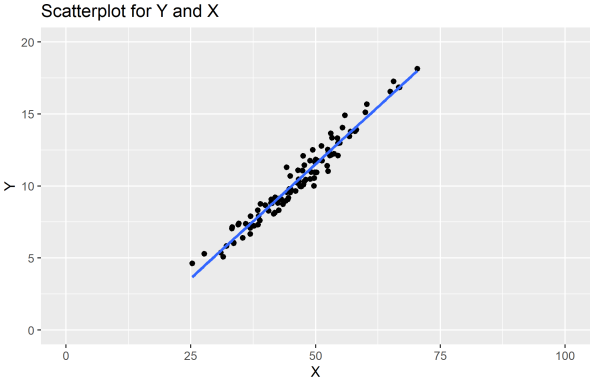
- Measurement error adjustments are tricky
  - Either require a 'gold standard' (a subset of our sample for which X is observed)
  - Or to rely on additional assumptions
- But often we can still anticipate its potential effects
  - If $E(\epsilon) \neq 0$ we should expect bias in the direction of the measurement error

    E.g. crack down policies on knife crime should be considered when assessing trends in knife crime using police data
  - If the measurement error is random and affecting the outcome variable, $E(Y^*) = Y$, only measures of uncertainty will be affected, $Y^* = \beta_0 + \beta_1 X + e + \epsilon$
  - However, even random error in an explanatory variable, will bias (attenuate) regression coefficients

    The slope in simple linear regression, $\hat{\beta}_1 = \dfrac{Cov(Y, X)}{Var(X)}$

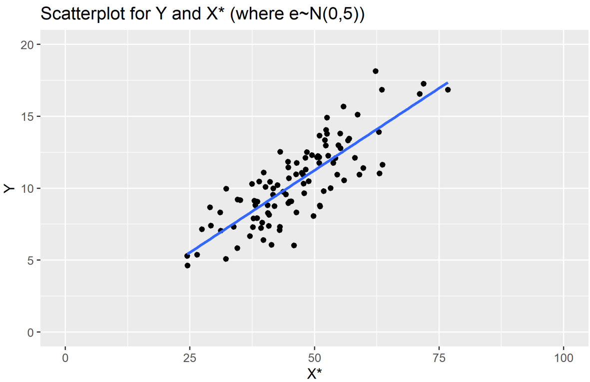
    If X is affected by random error, $\hat{\beta}_1^* = \dfrac{Cov(Y, X)}{Var(X) + Var(\epsilon)}$

Effect of Random Measurement Error

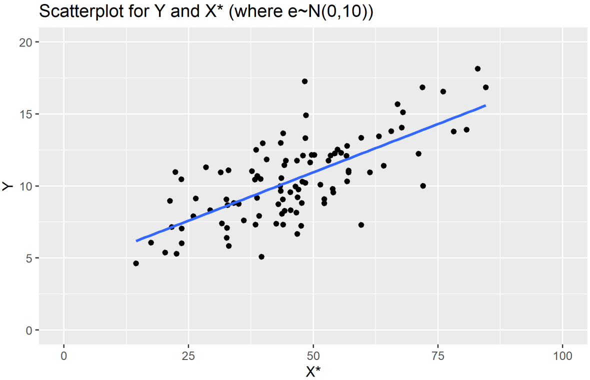


Workshop Aims
Selection Bias
Missing Data
Probability Weights
Imputation
Measurement Error
Recap

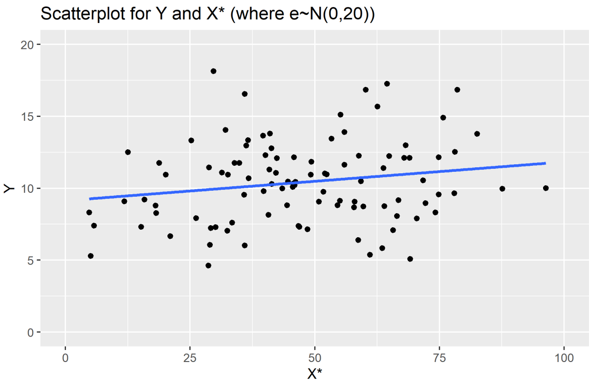
Effect of Random Measurement Error



Effect of Random Measurement Error



Effect of Random Measurement Error



Recap

- We have identified common consequences of missing data and measurement error
  - If the missing data is ignorable we should only expect a loss of statistical power
  - If the missing data is not ignorable we should expect bias
  - For measurement error even random error will bias our estimates (attenuate the slope)

Recap

- We have identified common consequences of missing data and measurement error
  - If the missing data is ignorable we should only expect a loss of statistical power
  - If the missing data is not ignorable we should expect bias
  - For measurement error even random error will bias our estimates (attenuate the slope)
- We have learnt some common methods to adjust these problems
  - Probability weights, help to improve overall representativity, easy to calculate and apply
  - Imputation, allow us to use cases affected by item-misingness

Recap

- We have identified common consequences of missing data and measurement error
  - If the missing data is ignorable we should only expect a loss of statistical power
  - If the missing data is not ignorable we should expect bias
  - For measurement error even random error will bias our estimates (attenuate the slope)
- We have learnt some common methods to adjust these problems
  - Probability weights, help to improve overall representativity, easy to calculate and apply
  - Imputation, allow us to use cases affected by item-misingness
- Recommended readings:
  - On probability weights Yansaneh (2003) 'Construction and Use of Sample Weights'
  - On multiple imputation Van Buuren & Groothuis-Oudshoorn (2013) 'mice: Multivariate Imputation by Chained Equations in R'