Quantitative Social Research II
Workshop 7: 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

- Assumptions in the linear regression model ($Y = \alpha + \beta_k X_k + e$):
  - normality: residuals are normally distributed
  - homoskedasticity: the variance of the residuals is constant
  - independence: residuals are independent of each other
  - no multicollinearity
  - **perfectly measured variables**
  - **no missing data** (other than missing at random)
  - no unobserved confounders: we control for all common causes of X_1 and Y
  - no reverse causality: Y does not cause X_1
  - linearity: the effect of X_1 on Y is the same across the range of X_1

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 (an entire case is missing)
  - item non-response (information for a given variable is missing)

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 (an entire case is missing)
  - item non-response (information for a given variable is missing)
- Missing data mechanisms can be classified in three groups
  - missing completely at random (MCAR) - not data dependent
  - missing at random (MAR) - dependent on seen data
  - missing not at random (MNAR) - dependent on unseen data
- Different implications depending on the ignorability of the missing data mechanism

UNIVERSITY OF LEEDS

Unit and Item Non-Response

- Question: can you identify which cases are affected by unit-missingness and which by item-missingness?

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 explanatory variables
  - e.g. some of the data was lost by accident
  - implications: loss of statistical power because of using a smaller sample

Missing Data Mechanisms

- Missing completely at random
  - the missing data mechanism is not related to any of our explanatory variables
  - e.g. some of the data was lost by accident
  - implications: loss of statistical power because of using a smaller sample
- Missing data at random
  - related to one or more of our explanatory variables
  - e.g. male judges might forget to submit their survey forms more commonly than female judges
  - 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 explanatory variables
  - e.g. some of the data was lost by accident
  - implications: loss of statistical power because of using a smaller sample
- Missing data at random
  - related to one or more of our explanatory variables
  - e.g. male judges might forget to submit their survey forms more commonly than female judges
  - if left unadjusted will bias our estimates, if adjusted becomes 'ignorable'
- Missing data not at random
  - systematically related to unobserved data
  - e.g. harsher judges might try to avoid submitting their forms
  - cannot be adjusted easily, will bias our estimates

Probability Weights

- Using *weights* we can reflect the over/under-representation of certain cases in our sample and obtain a more representative sample

Probability Weights

- Using *weights* we can reflect the over/under-representation of certain cases in our sample and obtain a more representative sample
- 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 over/under-representation of certain cases in our sample and obtain a more representative sample
- 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 can range from 0 to ∞, although in practice we often cap them from 0.3 to 3
  - 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 (validity) and precision (reliability)
- Can adjust for multiple variables
  - by combining their categories

    e.g. male-white, male-nonwhite, female-white, female-nonwhite
  - however, soon we run out of cases within specific categories
- Not so flexible to deal with item non-response
  - imputation methods are normally used instead

Making the most of the Data

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

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

Making the most of the Data

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

- Cases affected by unit-missing are dropped (case 4)
- But also we have to drop cases affected by item-missingness (cases 1 and 5) if we are using those variables, *listwise deletion*

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

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

- Cases affected by unit-missing are dropped (case 4)
- But also we have to drop cases affected by item-missingness (cases 1 and 5) if we are using those variables, *listwise deletion*

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

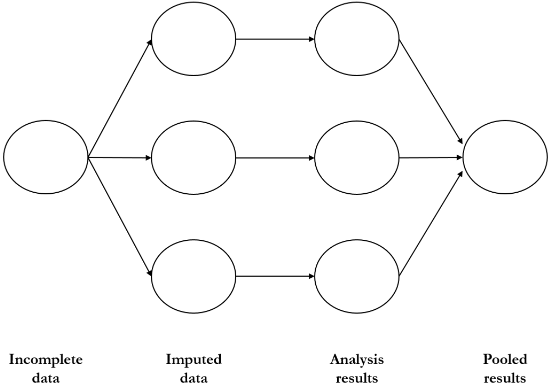
Single Imputation

- 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

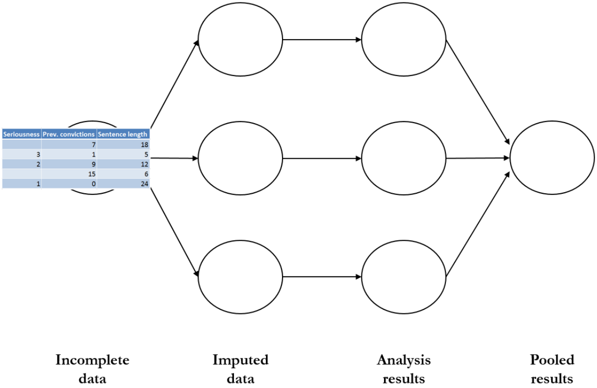
Multiple Imputation

- 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
  - downside, it is computationally intensive

Multiple Imputation

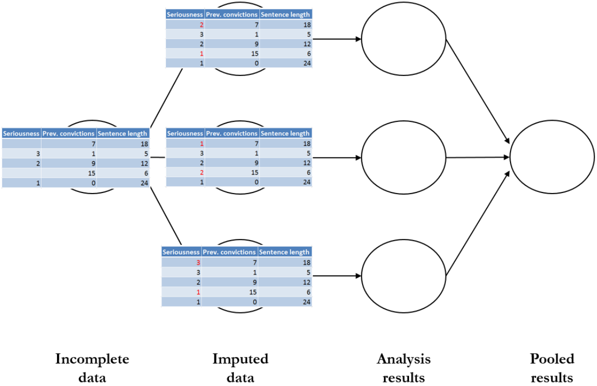


Multiple Imputation



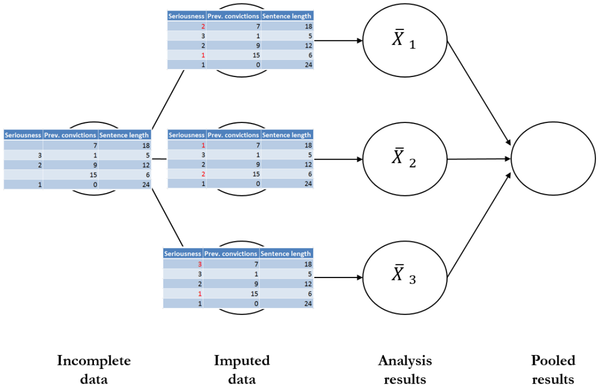
Seriousness	Prev. convictions	Sentence length
	7	18
3	1	5
2	9	12
	15	6
1	0	24

Multiple Imputation

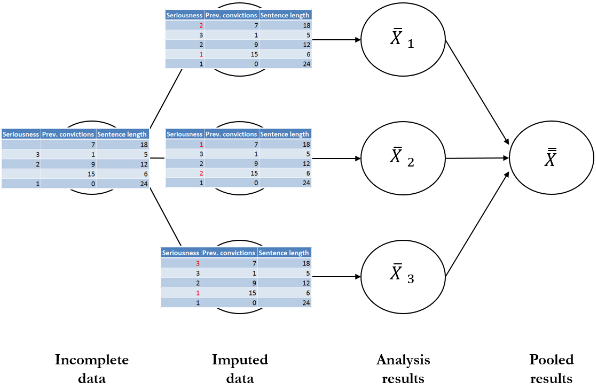


Workshop Aims
Selection Bias
Missing Data
Probability Weights
Imputation
Measurement Error
Recap

Multiple Imputation



Multiple Imputation



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

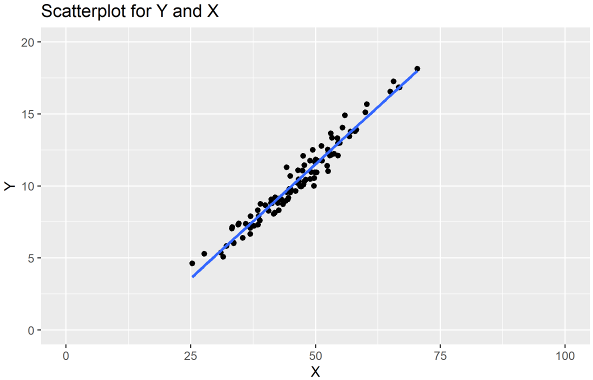
    e.g. crack down policies on knife crime should be considered when assessing trends in knife crime using police data
  - if the measurement error is random and affecting the outcome variable, $E(Y^*) = Y$, only measures of uncertainty will be affected, $Y^* = \beta_0 + \beta_1 X + e + \epsilon$
  - however,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

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

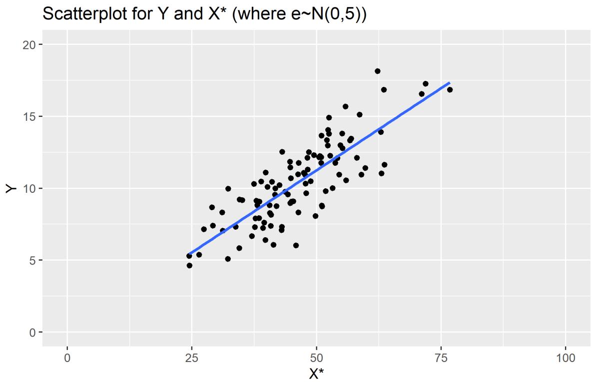
    if X is affected by random error, $\hat{\beta}_1^* = \dfrac{Cov(Y, X)}{Var(X) + Var(\epsilon)}$

UNIVERSITY OF LEEDS

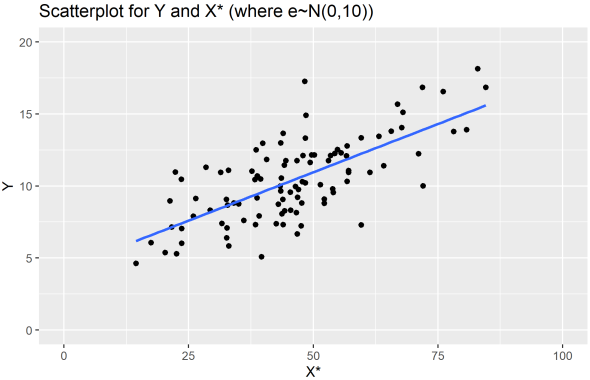
Effect of Random Measurement Error



Effect of Random Measurement Error

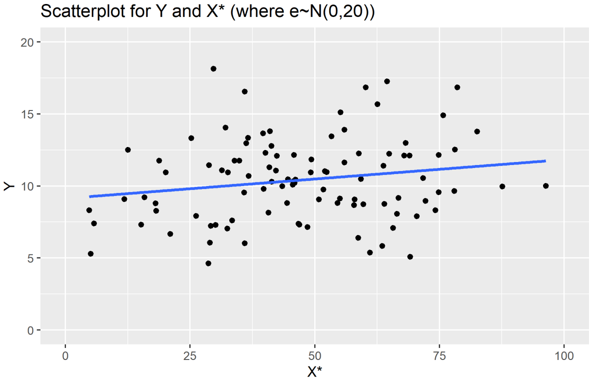


Effect of Random Measurement Error



UNIVERSITY OF LEEDS

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'