Introduction to Measurement Error and Misclassification

Jose Pina-Sánchez

Albert Varela

Introduction

- What is measurement error?
  - Discrepancies between the ‘true’ and the observed value
  - The result of poorly defined construct and/or an imperfect measurement process

Introduction

- What is measurement error?
  - Discrepancies between the 'true' and the observed value
  - The result of poorly defined construct and/or an imperfect measurement process
- Examples in the Social Sciences
  - Elusive constructs, loosely defined:
    e.g. happiness, ethnicity, political decentralisation
  - Subjectively elicited data:
    e.g. survey data, affected by memory failures (*when was the last time you went to a pub?*), social desirability (*for how long have you been unemployed?*), or content analysis affected by inter-rater unreliability
  - Administrative/official data used as proxies:
    e.g. using earnings to measure poverty, or measuring violent crime from police records
- Question: Can you think of examples form your research area?

Introduction

- Why does it matter?
  - We cannot describe reality accurately

    e.g. What is the true prevalence of covid? What is the true extent of property crime? Has it increased compared to last year? Is it higher in Leeds than in Bradford?
  - But also, our causal inferences will be biased

    e.g. Does education affect violent crime? To what extent a guilty plea lowers the probability of receiving a custodial sentence?

Introduction

- Why does it matter?
  - We cannot describe reality accurately

    e.g. What is the true prevalence of covid? What is the true extent of property crime? Has it increased compared to last year? Is it higher in Leeds than in Bradford?
  - But also, our causal inferences will be biased

    e.g. Does education affect violent crime? To what extent a guilty plea lowers the probability of receiving a custodial sentence?
- There are ways to anticipate the prevalence and impact of measurement error
  - And to some extent adjust for it
  - But to do so we first need to define these errors formally using measurement error models

Defining Measurement Error Formally

- The classical measurement error model (random errors)
  - $\overbrace{X^*}^{observed} = \overbrace{X}^{true\ value} + \overbrace{U}^{noise}$
  - with the errors taken to be randomly distributed, $U \sim N(0, \sigma_U)$

  - Question: Can you think of any examples?

Defining Measurement Error Formally

- The classical measurement error model (random errors)
  - $\overbrace{X^*}^{observed} = \overbrace{X}^{true\ value} + \overbrace{U}^{noise}$
  - with the errors taken to be randomly distributed, $U \sim N(0, \sigma_U)$

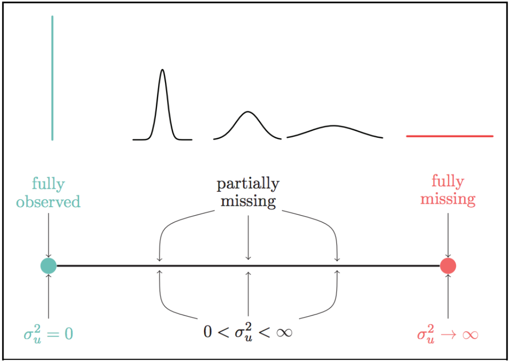
  - Question: Can you think of any examples?
    E.g. Results from a math test, blood pressure readings

Defining Measurement Error Formally

- The classical measurement error model (random errors)
  - $\overbrace{X^*}^{observed} = \overbrace{X}^{true\ value} + \overbrace{U}^{noise}$
  - with the errors taken to be randomly distributed, $U \sim N(0, \sigma_U)$

  - Question: Can you think of any examples?
    E.g. Results from a math test, blood pressure readings
- Only the variance is affected
  - $\sigma^2_{X^*} = \sigma^2_X + \sigma^2_U$; but the mean is unaffected since $E(U) = 0$
  - Taking repeated observations we can estimate the prevalence of classical measurement error
  - The reliability ratio: $\rho_{X^*} = \dfrac{\sigma^2_X}{\sigma^2_X + \sigma^2_U} = \dfrac{\text{true variability}}{\text{observed variability}}$

Conceptualising Random Errors



Source: Blackwell et al. (2017)

Systematic Errors

- The classical model is the most commonly used in applications seeking to describe and adjust for measurement error
  - It is simple, and reflects well enough some measurement error mechanisms, but not always

Systematic Errors

- The classical model is the most commonly used in applications seeking to describe and adjust for measurement error
  - It is simple, and reflects well enough some measurement error mechanisms, but not always
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$

  - Question: Can you think of any examples?

Systematic Errors

- The classical model is the most commonly used in applications seeking to describe and adjust for measurement error
  - It is simple, and reflects well enough some measurement error mechanisms, but not always
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$

  - Question: Can you think of any examples?
    E.g. crime recorded by the police, self-reported unemployment spells

Systematic Errors

- The classical model is the most commonly used in applications seeking to describe and adjust for measurement error
  - It is simple, and reflects well enough some measurement error mechanisms, but not always
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$

  - Question: Can you think of any examples?
    E.g. crime recorded by the police, self-reported unemployment spells
- Repeated observations won't pick up systematic errors
  - We need a *gold standard* (at least for a subgroup of our sample)
  - E.g. Unemployment register, victimisation surveys

Multiplicative Errors

- What if the error is proportional to the true value of the quantity being measured?
  - E.g. memory failures;

    How many alcohol units do you drink per week?

    How many times have you eaten out during the summer holidays?

    How many sex partners have you had in your life?

Multiplicative Errors

- What if the error is proportional to the true value of the quantity being measured?
  - E.g. memory failures;

    How many alcohol units do you drink per week?

    How many times have you eaten out during the summer holidays?

    How many sex partners have you had in your life?
- These can be better specified using a multiplicative rather than an additive model
  - I.e., as $X^* = X \cdot U$, rather than $X^* = X + U$,

    with $E(U) = 1$ if random, and $E(U) \neq 1$ if systematic

Misclassification

- Up to now we have taken examples from continuous variables
- What if the variable affected by measurement error is discrete?
  - E.g. self-reported work status, ethnicity determined through subjects' names
- We have a misclassification problem
  - which, for a binary variable, can be specified through a 2x2 misclassification matrix

$$\begin{cases} P(X^* = 1|X = 1) = \theta_{1|1}; & \text{Sensitivity} \\ P(X^* = 0|X = 0) = \theta_{0|0}; & \text{Specificity} \end{cases}$$

$$\begin{cases} P(X^* = 1|X = 0) = \theta_{1|0}; & \text{Probability false positive} \\ P(X^* = 0|X = 1) = \theta_{0|1}; & \text{Probability false negative} \end{cases}$$

Other Types of Errors

- We have reviewed the most common forms of measurement error, but there are many more that might be relevant

Other Types of Errors

- We have reviewed the most common forms of measurement error, but there are many more that might be relevant
- Heteroskedastic errors (their variance is not constant)
  - E.g. recall errors that increase with the age of the subject

Other Types of Errors

- We have reviewed the most common forms of measurement error, but there are many more that might be relevant
- Heteroskedastic errors (their variance is not constant)
  - E.g. recall errors that increase with the age of the subject
- Autocorrelated errors (they are correlated with each other)
  - Could be seen in hierarchical or longitudinal data; e.g. interviewer effects

Other Types of Errors

- We have reviewed the most common forms of measurement error, but there are many more that might be relevant
- Heteroskedastic errors (their variance is not constant)
  - E.g. recall errors that increase with the age of the subject
- Autocorrelated errors (they are correlated with each other)
  - Could be seen in hierarchical or longitudinal data; e.g. interviewer effects
- <u>Differential errors</u> (errors correlated with the residuals form the outcome model)
  - E.g. in studying the effect of employment status on self-esteem we should consider that the probability of misreporting employment status is correlated with self-esteem

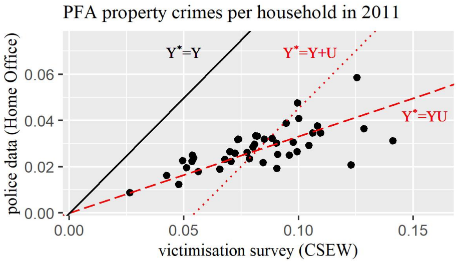
Multiple Error Mechanisms

- Often we see multiple measurement error mechanisms affecting the same variable
- This is what we found in the Recounting Crime project
  - Measurement error in police recorded crime rates

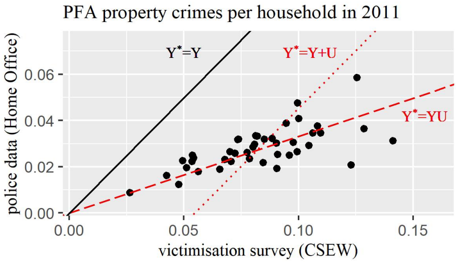
Multiple Error Mechanisms

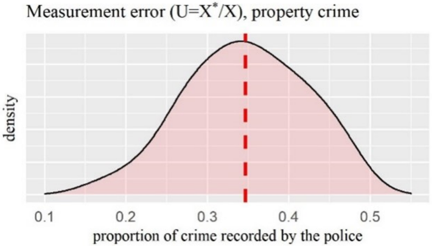
- Often we see multiple measurement error mechanisms affecting the same variable
- This is what we found in the Recounting Crime project
  - Measurement error in police recorded crime rates
- We defined this measurement error as:
  - *systematic*, since not all crime is reported to the police
  - *random*, subject to variability across areas, as a result of the different recording practices across police forces
  - *multiplicative*, errors seem proportional to the true extent of crime in the area

Multiplicative Errors: Crime Rates



Multiplicative Errors: Crime Rates





The Impact of Measurement Error

The Impact of Measurement Error

- We have seen how different forms of measurement error can affect univariate stats
  - Random errors affect measures of dispersion, systematic errors affect measures of centrality
- But how do (non-differential) errors affect estimates from multivariate (regression) models?

The Impact of Measurement Error

- We have seen how different forms of measurement error can affect univariate stats
  - Random errors affect measures of dispersion, systematic errors affect measures of centrality
- But how do (non-differential) errors affect estimates from multivariate (regression) models?
- Assuming only one variable is prone to measurement error, its impact will depend on:
  1. the outcome model (whether linear, Poisson, etc.)
  2. the measurement error model (additive, random, etc.)
  3. where is the affected variable introduced in the model (as a response or an explanatory variable)

Impact of Measurement Error

- Let's review some scenarios for the case of simple linear regression
  - $Y = \alpha + \beta X + \epsilon$

Impact of Measurement Error

- Let's review some scenarios for the case of simple linear regression
  - $Y = \alpha + \beta X + \epsilon$

1. Random additive errors affecting the response variable
   - $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$
   - E.g. IQ scores, measures of blood pressure

Impact of Measurement Error

- Let's review some scenarios for the case of simple linear regression
  - $Y = \alpha + \beta X + \epsilon$

1. Random additive errors affecting the response variable
   - $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$
   - E.g. IQ scores, measures of blood pressure
2. Similar errors affecting the explanatory variable
   - $X^* = X + U$, and $U \sim N(0, \sigma_U)$

Impact of Measurement Error

- Let's review some scenarios for the case of simple linear regression
  - $Y = \alpha + \beta X + \epsilon$

1. Random additive errors affecting the response variable
   - $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$
   - E.g. IQ scores, measures of blood pressure
2. Similar errors affecting the explanatory variable
   - $X^* = X + U$, and $U \sim N(0, \sigma_U)$
3. Systematic additive errors affecting the response variable
   - $Y^* = Y + U$, and $E(U) \neq 0$
   - E.g. self-reported position in a scale of xenophilia, percentage of income donated to charities

Impact of Measurement Error

- Let's review some scenarios for the case of simple linear regression
  - $Y = \alpha + \beta X + \epsilon$

1. Random additive errors affecting the response variable
   - $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$
   - E.g. IQ scores, measures of blood pressure
2. Similar errors affecting the explanatory variable
   - $X^* = X + U$, and $U \sim N(0, \sigma_U)$
3. Systematic additive errors affecting the response variable
   - $Y^* = Y + U$, and $E(U) \neq 0$
   - E.g. self-reported position in a scale of xenophilia, percentage of income donated to charities
4. Systematic multiplicative errors affecting the response variable
   - $Y^* = Y \cdot U$, and $E(U) \neq 1$
   - E.g. self-reported duration of spells of unemployment, police recorded crime across areas

Impact of Measurement Error

- Let's review some scenarios for the case of simple linear regression
  - $Y = \alpha + \beta X + \epsilon$

1. Random additive errors affecting the response variable
   - $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$
   - E.g. IQ scores, measures of blood pressure
2. Similar errors affecting the explanatory variable
   - $X^* = X + U$, and $U \sim N(0, \sigma_U)$
3. Systematic additive errors affecting the response variable
   - $Y^* = Y + U$, and $E(U) \neq 0$
   - E.g. self-reported position in a scale of xenophilia, percentage of income donated to charities
4. Systematic multiplicative errors affecting the response variable
   - $Y^* = Y \cdot U$, and $E(U) \neq 1$
   - E.g. self-reported duration of spells of unemployment, police recorded crime across areas

- Question: Will β be biased in any of those scenarios?

Classical Error on the Response

- Scenario 1: random additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$

Classical Error on the Response

- Scenario 1: random additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$

    $Y + U = \alpha + \beta X + \epsilon$

Classical Error on the Response

- Scenario 1: random additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$

    $Y + U = \alpha + \beta X + \epsilon$

    $Y = \alpha + \beta X + (\epsilon - U)$
  - The measurement error is absorbed by the model's error term, affecting precision, but leaving regression coefficients unbiased

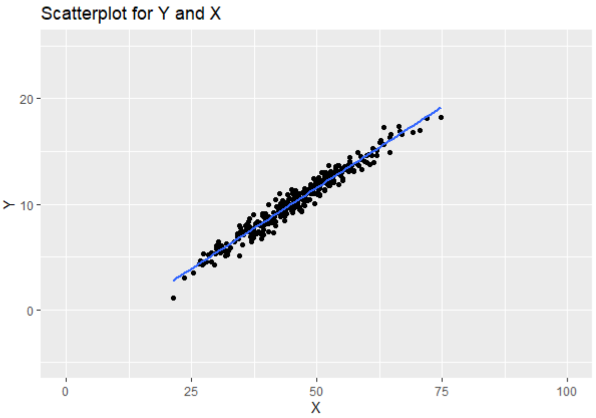
Classical Error on the Response

- Scenario 1: random additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $U \sim N(0, \sigma_U)$

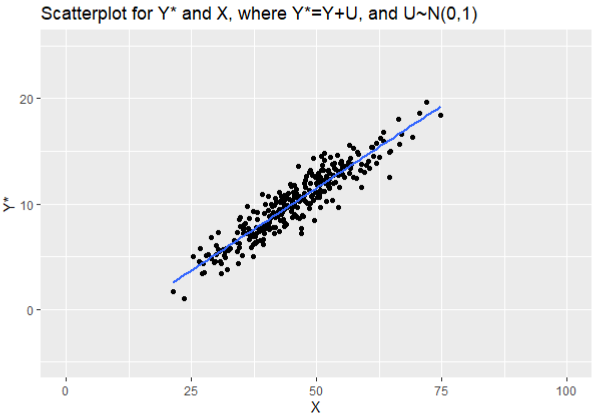
    $Y + U = \alpha + \beta X + \epsilon$

    $Y = \alpha + \beta X + (\epsilon - U)$
  - The measurement error is absorbed by the model's error term, affecting precision, but leaving regression coefficients unbiased
  - We can see this effect using simulated data

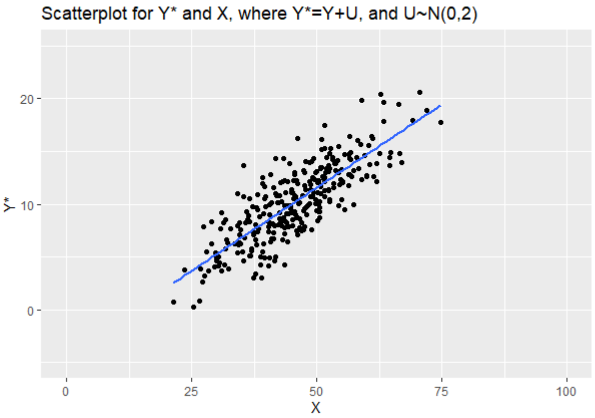
Classical Error on the Response



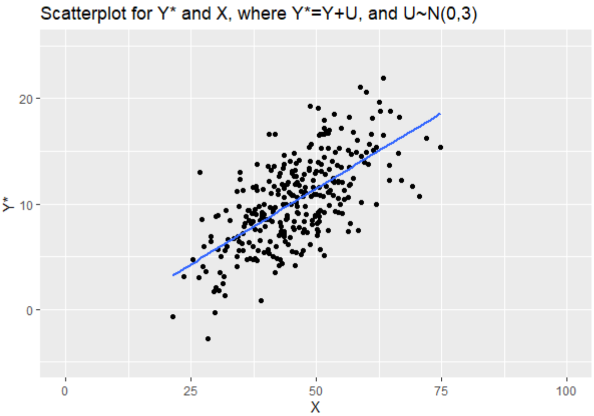
Classical Error on the Response



Classical Error on the Response



Classical Error on the Response



Classical Error on a Covariate

- Scenario 2: random additive errors on the covariate
  - $Y = \alpha + \beta X^* + \epsilon$, with $X^* = X + U$, and $U \sim N(0, \sigma_U)$

Classical Error on a Covariate

- Scenario 2: random additive errors on the covariate
  - $Y = \alpha + \beta X^* + \epsilon$, with $X^* = X + U$, and $U \sim N(0, \sigma_U)$
  - Using OLS we can estimate α and β solving...

$$\begin{cases} \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \\ \widehat{\beta} = \dfrac{\sigma_{XY}}{\sigma_X^2} \end{cases}$$

Classical Error on a Covariate

- Scenario 2: random additive errors on the covariate
  - $Y = \alpha + \beta X^* + \epsilon$, with $X^* = X + U$, and $U \sim N(0, \sigma_U)$
  - Using OLS we can estimate α and β solving...

$$\begin{cases} \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \\ \widehat{\beta} = \dfrac{\sigma_{XY}}{\sigma_X^2} \end{cases}$$

  - If instead we have...

$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*Y}}{\sigma_{X^*}^2} \end{cases}$$

Classical Error on a Covariate

- Scenario 2: random additive errors on the covariate
  - $Y = \alpha + \beta X^* + \epsilon$, with $X^* = X + U$, and $U \sim N(0, \sigma_U)$
  - Using OLS we can estimate α and β solving...

$$\begin{cases} \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \\ \widehat{\beta} = \dfrac{\sigma_{XY}}{\sigma_X^2} \end{cases}$$

  - If instead we have...then..

$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* = \bar{Y} - \widehat{\beta}\bar{X} = \widehat{\alpha}; \quad \underline{\text{unbiased constant}} \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*Y}}{\sigma_{X^*}^2} \end{cases}$$

Classical Error on a Covariate

- Scenario 2: random additive errors on the covariate
  - $Y = \alpha + \beta X^* + \epsilon$, with $X^* = X + U$, and $U \sim N(0, \sigma_U)$
  - Using OLS we can estimate α and β solving...

$$\begin{cases} \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \\ \widehat{\beta} = \dfrac{\sigma_{XY}}{\sigma^2_X} \end{cases}$$

  - If instead we have X^*, then..

$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* = \bar{Y} - \widehat{\beta}\bar{X} = \widehat{\alpha}; & \underline{\text{unbiased constant}} \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*,Y}}{\sigma^2_{X^*}} = \dfrac{\sigma_{X,Y}}{\sigma^2_X + \sigma^2_U} = \widehat{\beta}\left(\dfrac{\sigma^2_X}{\sigma^2_X + \sigma^2_U}\right); & \underline{\text{attenuated slope}} \end{cases}$$

Classical Error on a Covariate

- Scenario 2: random additive errors on the covariate
  - $Y = \alpha + \beta X^* + \epsilon$, with $X^* = X + U$, and $U \sim N(0, \sigma_U)$
  - Using OLS we can estimate α and β solving...

$$\begin{cases} \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \\ \widehat{\beta} = \dfrac{\sigma_{XY}}{\sigma^2_X} \end{cases}$$

  - If instead we have X^*, then..

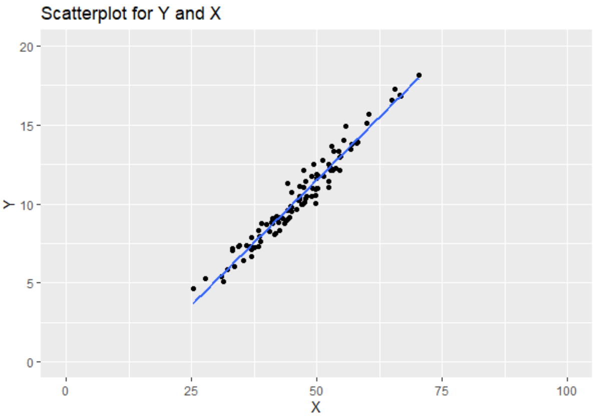
$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* = \bar{Y} - \widehat{\beta}\bar{X} = \widehat{\alpha}; & \underline{\text{unbiased constant}} \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*,Y}}{\sigma^2_{X^*}} = \dfrac{\sigma_{X,Y}}{\sigma^2_X + \sigma^2_U} = \widehat{\beta}\left(\dfrac{\sigma^2_X}{\sigma^2_X + \sigma^2_U}\right); & \underline{\text{attenuated slope}} \end{cases}$$

Using the properties of the covariance and the variance we can see how the former is not affected by random noise, but the latter is:

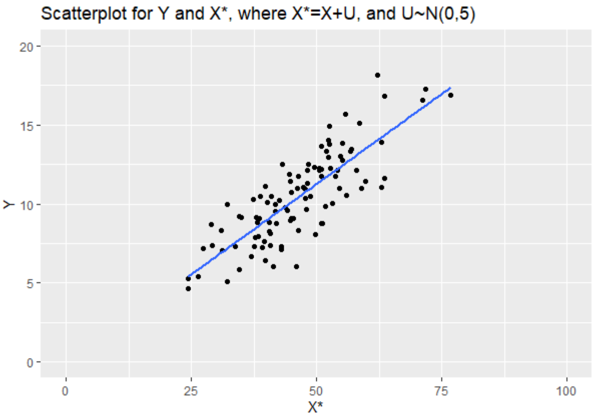
$$\sigma_{X^*,Y} = \sigma_{X+U,Y} = \sigma_{X,Y} + \sigma_{U,Y} = \sigma_{X,Y}$$
$$\sigma^2_{X^*} = \sigma^2_{X+U} = \sigma^2_X + \sigma^2_U + \sigma_{X,U} = \sigma^2_X + \sigma^2_U$$

  - We can see this effect using simulated data

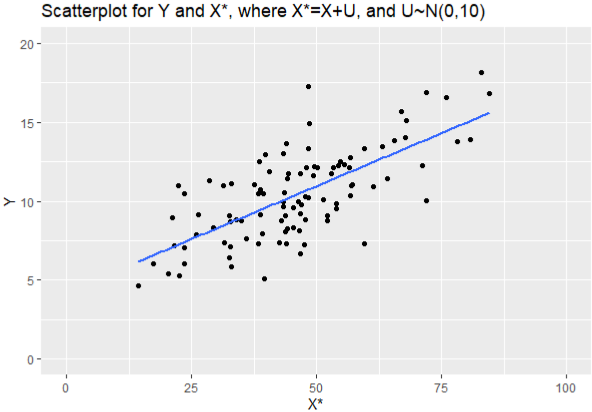
Effect of Random Measurement Error



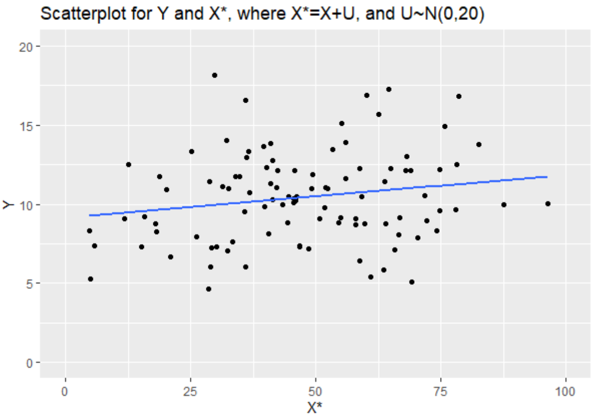
Effect of Random Measurement Error



Effect of Random Measurement Error



Effect of Random Measurement Error



Systematic Errors on the Response

- Scenario 3: systematic additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $\underline{E(U) \neq 0}$

Systematic Errors on the Response

- Scenario 3: systematic additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $\underline{E(U) \neq 0}$

    $Y + U = \alpha + \beta X + \epsilon$

    $Y = (\alpha - U) + \beta X + \epsilon$
  - The constant is biased, but the slope is not

Systematic Errors on the Response

- Scenario 3: systematic additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $\underline{E(U) \neq 0}$

    $Y + U = \alpha + \beta X + \epsilon$

    $Y = (\alpha - U) + \beta X + \epsilon$
  - The constant is biased, but the slope is not
- Scenario 4: systematic multiplicative errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $\underline{Y^* = Y \cdot U}$, and $E(U) \neq 1$

Systematic Errors on the Response

- Scenario 3: systematic additive errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $Y^* = Y + U$, and $\underline{E(U) \neq 0}$

    $Y + U = \alpha + \beta X + \epsilon$

    $Y = (\alpha - U) + \beta X + \epsilon$
  - The constant is biased, but the slope is not
- Scenario 4: systematic multiplicative errors on the response
  - $Y^* = \alpha + \beta X + \epsilon$, with $\underline{Y^* = Y \cdot U}$, and $E(U) \neq 1$

    $Y \cdot U = \alpha + \beta X + \epsilon$

    $$Y = \frac{\alpha + \beta X + \epsilon}{U}$$
  - All regression coefficients are biased

Impact of Measurement Error

- Depending on the type of errors, we see vastly different impacts
  - From relatively negligible to ‘all is wrong!’
  - Even when the errors are completely random

Impact of Measurement Error

- Depending on the type of errors, we see vastly different impacts
  - From relatively negligible to 'all is wrong!'
  - Even when the errors are completely random
- We have only considered simple linear regression
  - When we add other explanatory variables their coefficients will also be biased, even if they are perfectly measured

    Assume multiple linear regression, with one error-prone variable

    $Y = \alpha + \beta_1 X^* + \beta_2 Z + \epsilon$

    Both slopes will be biased; we will estimate $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ instead

    $$\widehat{\beta}_1^* = \widehat{\beta}_1 \left(\frac{\sigma^2_{X|Z}}{\sigma^2_{X|Z} + \sigma^2_U} \right) = \widehat{\beta}_1 \rho'_{X^*}$$

    $$\widehat{\beta}_2^* = \widehat{\beta}_2 + \widehat{\beta}_1 (1 - \rho'_{X^*}) \Gamma_Z$$

Impact of Measurement Error

- Depending on the type of errors, we see vastly different impacts
  - From relatively negligible to 'all is wrong!'
  - Even when the errors are completely random
- We have only considered simple linear regression
  - When we add other explanatory variables their coefficients will also be biased, even if they are perfectly measured

    Assume multiple linear regression, with one error-prone variable

    $Y = \alpha + \beta_1 X^* + \beta_2 Z + \epsilon$

    Both slopes will be biased; we will estimate $\widehat{\beta}_1^*$ and $\widehat{\beta}_2^*$ instead

    $$\widehat{\beta}_1^* = \widehat{\beta}_1 \left(\frac{\sigma^2_{X|Z}}{\sigma^2_{X|Z} + \sigma^2_U} \right) = \widehat{\beta}_1 \rho'_{X^*}$$

    $$\widehat{\beta}_2^* = \widehat{\beta}_2 + \widehat{\beta}_1 (1 - \rho'_{X^*}) \Gamma_Z$$
  - Much harder to assess if we consider more complex models: non-linear, two-stage processes, systems of equations, etc.
  - For most 'real-world' applications we won't be able to trace out the measurement error induced biased algebraically
  - And we have not even consider how measurements of uncertainty are also affected

Impact of Measurement Error

- In the words of Nugent et al. (2000: 60):
  - *"Measurement error is, to borrow a metaphor, a gremlin hiding in the details of our research that can contaminate the entire set of estimated regression parameters"*