Introduction to Measurement Error

Jose Pina-Sánchez
David Buil-Gil

Introduction

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

Introduction

- What is measurement error?
  - any discrepancy between the 'true' and the observed value
  - the result of poorly defined construct and/or an imperfect measurement process
- Why does it matter?
  - we cannot describe reality accurately

    e.g. What is the true extent of property crime? What is the true prevalence of covid? Has it increased compared to last year? Is it higher in Leeds than in Bradford?
  - measurement error can also bias causal inference

    e.g. Does education affect violent crime? Does crime affect property values?

Introduction

- What is measurement error?
  - any discrepancy between the ‘true’ and the observed value
  - the result of poorly defined construct and/or an imperfect measurement process
- Why does it matter?
  - we cannot describe reality accurately

    e.g. What is the true extent of property crime? What is the true prevalence of covid? Has it increased compared to last year? Is it higher in Leeds than in Bradford?
  - measurement error can also bias causal inference

    e.g. Does education affect violent crime? Does crime affect property values?
- There are ways to anticipate its impact
  - and to some extent adjust for it
  - but to do so we first need to define these errors formally

Defining Measurement Error Formally

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, $U \sim N(0, \sigma_U)$

  - 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, $U \sim N(0, \sigma_U)$

  - 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: $\rho_{X^*} = \dfrac{\sigma^2_X}{\sigma^2_X + \sigma^2_U} = \dfrac{\text{true variability}}{\text{observed variability}}$

Systematic Errors

- The classical model is the most commonly invoked
  - it is simple and reflects well enough some measurement processes, but not all

Systematic Errors

- The classical model is the most commonly invoked
  - it is simple and reflects well enough some measurement processes, but not all
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$

  - e.g. self-reported spells of unemployment, police recorded crime

Systematic Errors

- The classical model is the most commonly invoked
  - it is simple and reflects well enough some measurement processes, but not all
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$

  - e.g. self-reported spells of unemployment, police recorded crime
- Repeated observations won't detect systematic errors
  - we need a *gold standard*
  - 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, count data

    How many alcohol units do you drink per week?

    How many 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, count data

    How many alcohol units do you drink per week?

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

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

Other Types of Errors

Other Types of Errors

- Misclassification (for discrete data)
  - e.g. ethnicity determined through individuals' names

Other Types of Errors

- Misclassification (for discrete data)
  - e.g. ethnicity determined through individuals' names
- Heteroskedastic errors (their variance is not constant)
  - e.g. recall errors that increase with the age of the subject

Other Types of Errors

- Misclassification (for discrete data)
  - e.g. ethnicity determined through individuals' names
- 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)
  - in spatial and longitudinal data; e.g. interviewer effects

Other Types of Errors

- Misclassification (for discrete data)
  - e.g. ethnicity determined through individuals' names
- 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)
  - in spatial and longitudinal data; e.g. interviewer effects
- Differential errors (correlated with the cause or the effect)

The Impact of Measurement Error

- Different forms of measurement error can affect univariate stats
  - random errors affect measures of dispersion, systematic errors affect measures of centrality

The Impact of Measurement Error

- Different forms of measurement error can affect univariate stats
  - random errors affect measures of dispersion, systematic errors affect measures of centrality
- But how does measurement error affect estimates from multivariate (regression) models?

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 $U \sim N(0, \sigma_U)$

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 $U \sim N(0, \sigma_U)$
2. Similar random 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 $U \sim N(0, \sigma_U)$
2. Similar random 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 $E(U) \neq 0$

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 $U \sim N(0, \sigma_U)$
2. Similar random 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 $E(U) \neq 0$
4. Systematic multiplicative errors affecting the response variable
   - $Y^* = Y \cdot U$, and $E(U) \neq 1$

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 $U \sim N(0, \sigma_U)$
2. Similar random 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 $E(U) \neq 0$
4. Systematic multiplicative errors affecting the response variable
   - $Y^* = Y \cdot U$, and $E(U) \neq 1$

- Question: Is β 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 $U \sim N(0, \sigma_U)$

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

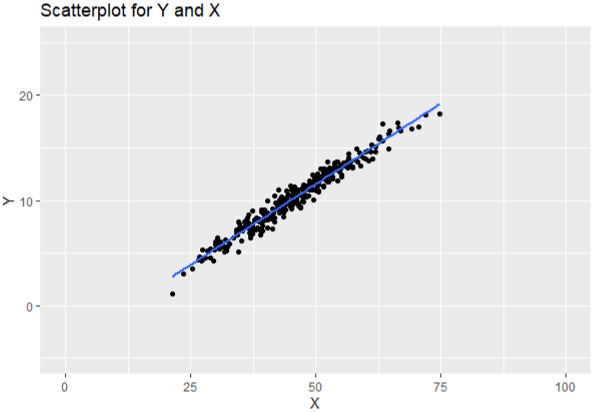
Classical Error on the Response

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

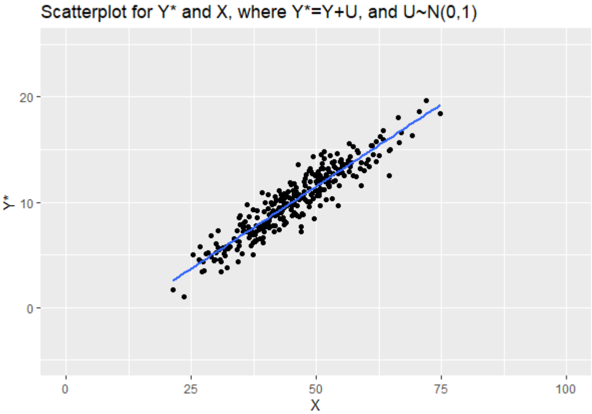
    $Y + U = \alpha + \beta X + \epsilon$

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

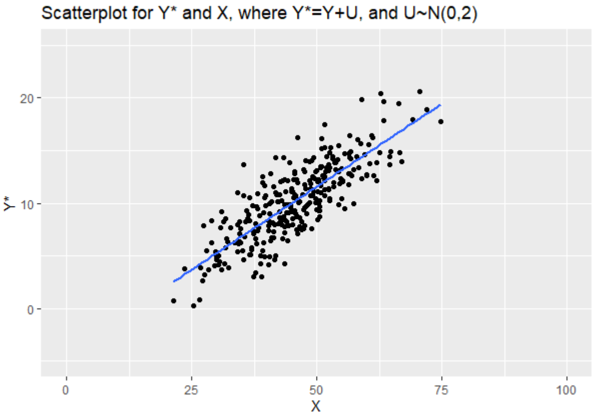
Classical Error on the Response



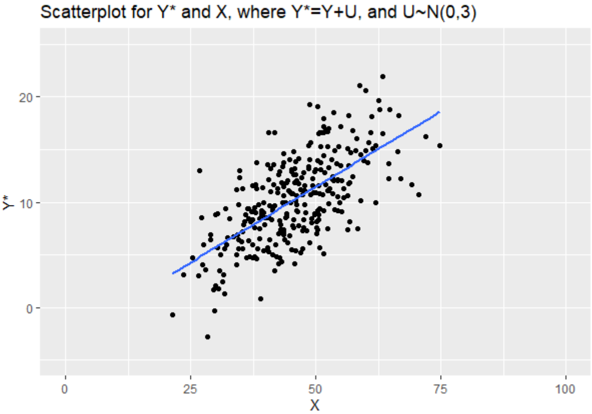
Classical Error on the Response



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...

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

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...

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

  - If instead we have...

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

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...

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

  - If instead we have...then..

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

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...

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

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

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

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...

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

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

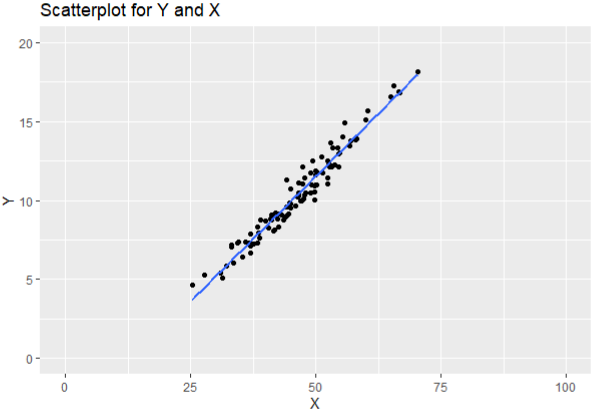
$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* = \bar{Y} - \widehat{\beta}\bar{X} = \widehat{\alpha}; & \underline{\text{unbiased constant}} \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*,Y}}{\sigma_{X^*}^2} = \dfrac{\sigma_{X,Y}}{\sigma_X^2 + \sigma_U^2} = \widehat{\beta}\left(\dfrac{\sigma_X^2}{\sigma_X^2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

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:

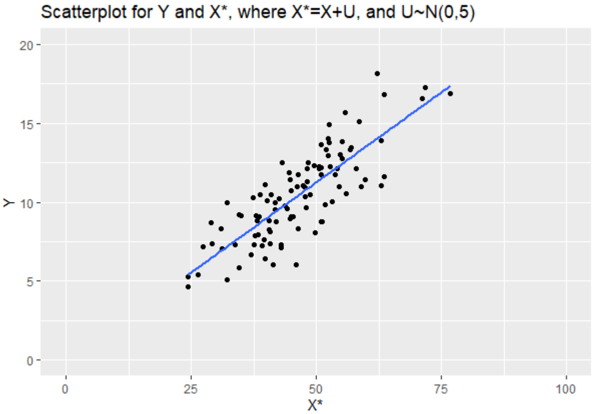
$$\sigma_{X^*,Y} = \sigma_{X+U,Y} = \sigma_{X,Y} + \sigma_{U,Y} = \sigma_{X,Y}$$
$$\sigma_{X^*}^2 = \sigma_{X+U}^2 = \sigma_X^2 + \sigma_U^2 + \sigma_{X,U} = \sigma_X^2 + \sigma_U^2$$

  - We can see this effect using simulated data

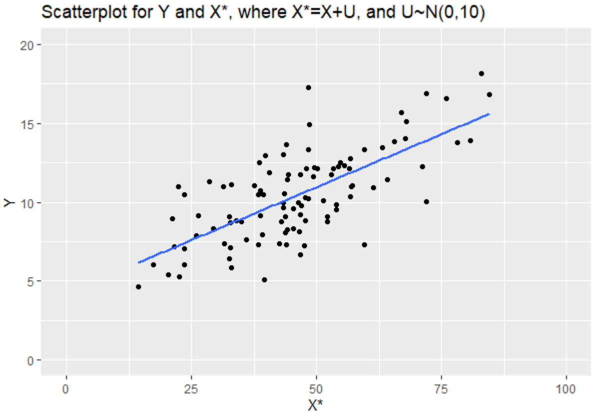
Effect of Random Measurement Error



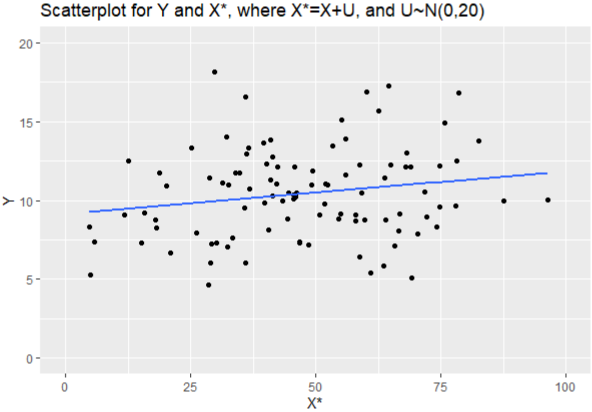
Effect of Random Measurement Error



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
- 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”*

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
- 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"*