Introduction to Measurement Error: Prevalence, Impact and Adjustments

Jose Pina-Sánchez (University of Leeds)
j.pinasanchez@leeds.ac.uk

2021 Research Methods e Festival
Hosted by NCRM and methods@manchester

Introduction

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

Introduction

- What is measurement error?
  - The result of an imperfect measurement process
  - Discrepancies between the 'true' and the observed value
- A widespread problem, but especially prevalent in the Social Sciences
  - Elusive constructs, subjectively defined, e.g. happiness, ethnicity
  - Even when well-defined, they often are subjectively elicited: e.g. survey data, affected by memory failures (e.g. *when was the last time you went to a pub?*), social desirability (e.g. *for how long have you been unemployed*?)
  - Administrative/official data used as proxies (e.g. using earnings to measure poverty; measuring violent crime from police records)

Introduction

- What is measurement error?
  - The result of an imperfect measurement process
  - Discrepancies between the 'true' and the observed value
- A widespread problem, but especially prevalent in the Social Sciences
  - Elusive constructs, subjectively defined, e.g. happiness, ethnicity
  - Even when well-defined, they often are subjectively elicited: e.g. survey data, affected by memory failures (e.g. *when was the last time you went to a pub?*), social desirability (e.g. *for how long have you been unemployed?*)
  - Administrative/official data used as proxies (e.g. using earnings to measure poverty; measuring violent crime from police records)
- To investigate the prevalence, impact and adjustments we 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, $U \sim N(0, \sigma_U)$
  - E.g. results from an IQ test, blood pressure measurements,

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 an IQ test, blood pressure measurements,

- 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 used in applications seeking to describe and adjust for measurement error
  - It is simple, and reflects well enough some measurement error mechanisms, but it is often misleading

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 it is often misleading
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$
  - E.g. crime reported to the police, self-reported xenophobia, whether received benefits

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 it is often misleading
- Measurement error is often *systematic*
  - $X^* = X + U$; but $E(U) \neq 0$
  - E.g. crime reported to the police, self-reported xenophobia, whether received benefits

- 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

More Errors (Multiplicative)?!

- 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 indoors (in a restaurant or similar) since 'Freedom day'?

More Errors (Multiplicative)?!

- 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 indoors (in a restaurant or similar) since 'Freedom day'?
  - 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

More Errors (Multiplicative)?!

- 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 indoors (in a restaurant or similar) since 'Freedom day'?
  - These can be better specified using a multiplicative rather than an additive model

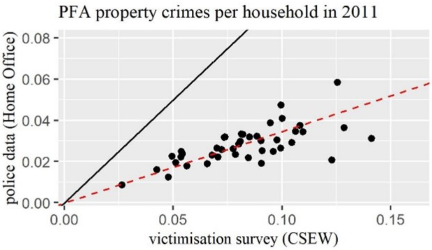
    I.e., as $X^* = X \cdot U$, rather than $X^* = X + U$,

    with $E(U) = 1$ if random, and $E(U) \neq 1$ if systematic
- Often we might have multiple mechanisms
  - Recounting Crime, an ESRC project where we study measurement error in police recorded crime rates
  - We have defined the measurement error as:

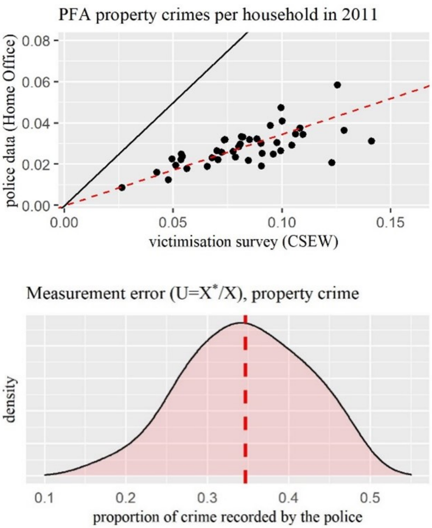
    i) multiplicative systematic, since not all crime is reported to the police, and the observed crime seems proportional to the true extent of crime in the area

    ii) subject to variability across areas, as a result of the different reporting rates and recording practices across police forces

Multiplicative Errors: Crime Rates



Multiplicative Errors: Crime Rates



More Errors (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

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

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

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 does measurement error 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 does measurement error 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

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

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

UNIVERSITY OF LEEDS

SDAI Secondary Data Analysis Initiative

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

methods@manchester

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

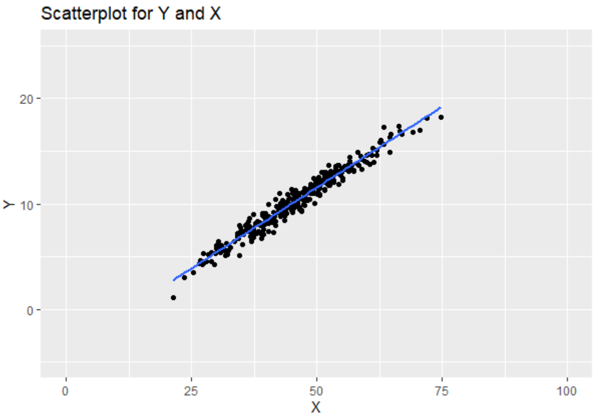
    $Y + U = \alpha + \beta X + \epsilon$

    $Y = \alpha + \beta X + (\epsilon - U)$
  - 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

    $Y \sim N(10, 3); \quad X = 15 + 3Y + \epsilon; \quad \rho_{Y,X} = 0.97$

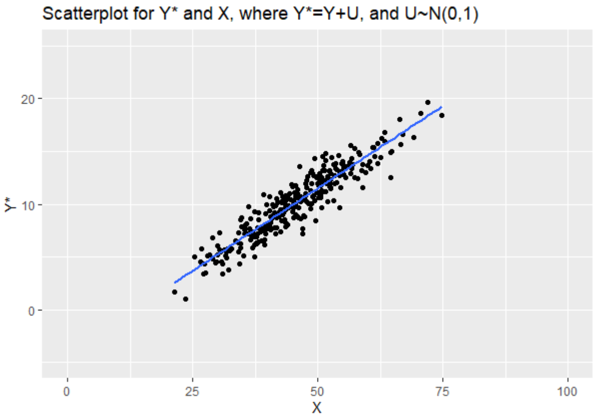
    $\sigma_U = (0, 1, 2, 3)$

Classical Error on the Response

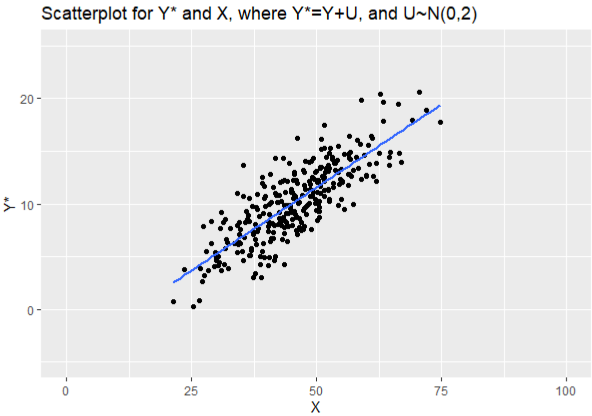


UNIVERSITY OF LEEDS

Classical Error on the Response

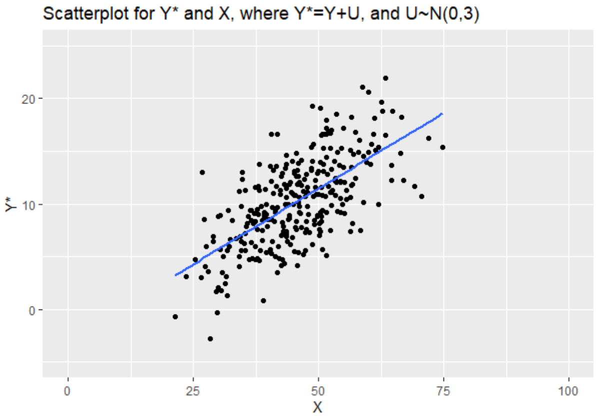


Classical Error on the Response



UNIVERSITY OF LEEDS

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...

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

  - If instead we have...then..

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

UNIVERSITY OF LEEDS

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..

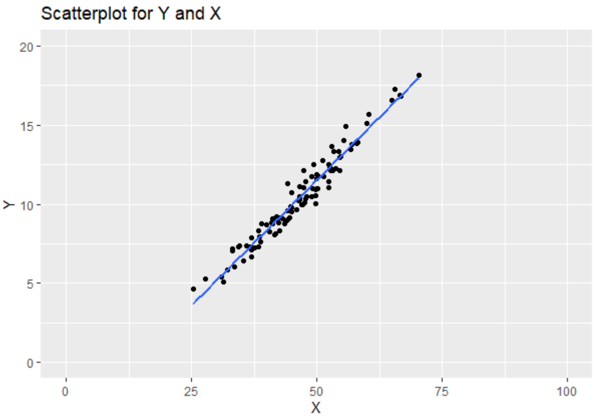
$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* = \bar{Y} - \widehat{\beta}\bar{X} = \widehat{\alpha}; & \underline{\text{unbiased constant}} \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*Y}}{\sigma_{X^*}^2} = \dfrac{\sigma_{XY}}{\sigma_X^2 + \sigma_U^2} = \widehat{\beta}\left(\dfrac{\sigma_X^2}{\sigma_X^2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

  - We can see this effect using simulated data

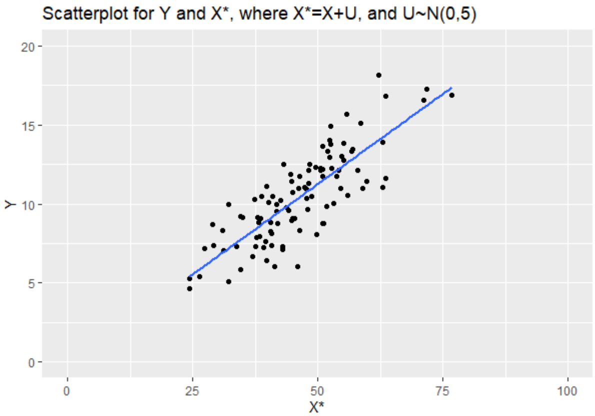
$Y \sim N(10, 3); \quad X = 15 + 3Y + \epsilon; \quad \rho_{Y,X} = 0.97$

$\sigma_U = (0, 5, 10, 20)$

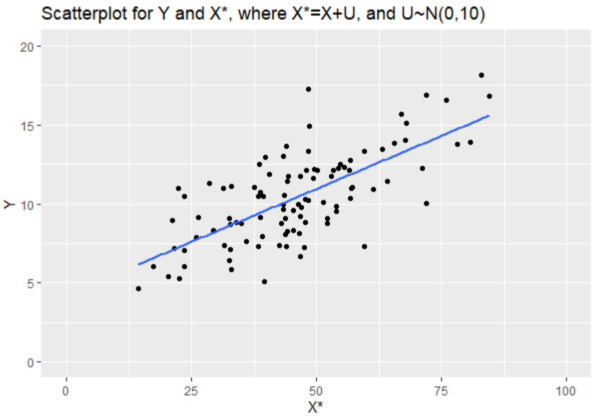
Effect of Random Measurement Error



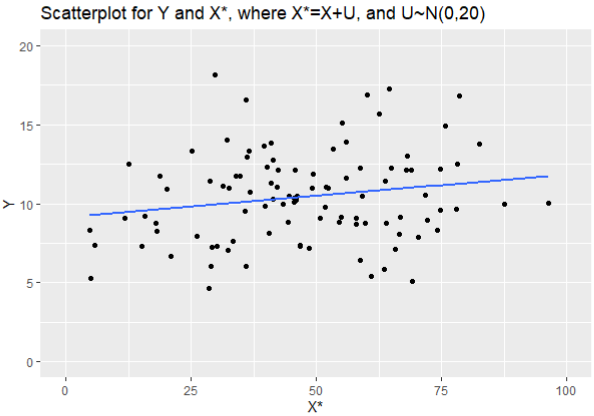
Effect of Random Measurement Error



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

    E.g. When modelling police crime rates ($E(U) < 1$), regression coefficients are biased upwards

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
  - 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

- 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
  - 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
- 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"*

Adjustment Methods

- We should always aim to improve data collection processes
- When that is not possible we should adjust for the impact of measurement error
  - We have seen how to do this in some simple settings, where we can anticipate the impact

Adjustment Methods

- We should always aim to improve data collection processes
- When that is not possible we should adjust for the impact of measurement error
  - We have seen how to do this in some simple settings, where we can anticipate the impact
- Ex.1, the effect of perceptions of offender dangerousness (subject to classical errors) on the duration of prison sentences
  - the measurement error variability can be derived from the inter-rater reliability for a subsample of cases,
  - which can then be used to adjust the expected bias

$$\widehat{\beta}^* = \widehat{\beta}\left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\right)$$

Adjustment Methods

- We should always aim to improve data collection processes
- When that is not possible we should adjust for the impact of measurement error
  - We have seen how to do this in some simple settings, where we can anticipate the impact
- Ex.1, the effect of perceptions of offender dangerousness (subject to classical errors) on the duration of prison sentences
  - the measurement error variability can be derived from the inter-rater reliability for a subsample of cases,
  - which can then be used to adjust the expected bias
    $$\widehat{\beta}^* = \widehat{\beta}\left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}\right)$$
- Ex.2, the effect of immigration on crime rates recorded by the police (systematic multiplicative errors)
  - the under-recording can be estimated using victimisation surveys,
    $\widehat{\beta}^* = \widehat{\beta}/\bar{U}$

UNIVERSITY OF LEEDS

Adjustment Methods

- When we can't trace out the impact of measurement error algebraically we need to use other methods

Adjustment Methods

- When we can't trace out the impact of measurement error algebraically we need to use other methods
- Most adjustment methods require additional forms of data
  - Multiple reflective indicators (latent variable models)
  - Instrumental variables (two stage processes)
  - A validation subsample (multiple imputation)
  - Repeated observations (regression calibration)

Adjustment Methods

- When we can't trace out the impact of measurement error algebraically we need to use other methods
- Most adjustment methods require additional forms of data
  - Multiple reflective indicators (latent variable models)
  - Instrumental variables (two stage processes)
  - A validation subsample (multiple imputation)
  - Repeated observations (regression calibration)
- Others can be used when all you have is an educated guess (sensitivity analysis)
  - Bayesian adjustments (Gustaffson, 2003)
  - Multiple overimputation (Blackwell et al., 2017)
  - SIMEX (Cook & Stefanski, 1994)

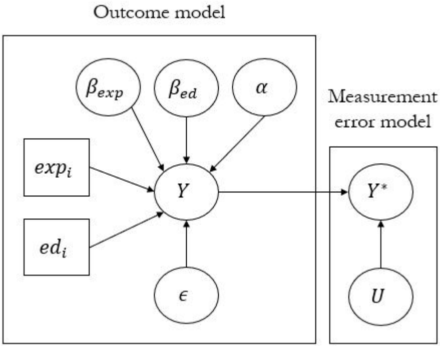
Bayesian Adjustments

- The most flexible approach
  - Can adjust any form of measurement error
  - We just need to specify a measurement model to be estimated simultaneously with our outcome model
  - The more we know about the underlying measurement error mechanisms the better the measurement model can be specified
  - The better the measurement model the more effective the adjustment
  - Even if we have an educated guess, it is still worth doing as a sensitivity analysis

Bayesian Adjustments

- The most flexible approach
  - Can adjust any form of measurement error
  - We just need to specify a measurement model to be estimated simultaneously with our outcome model
  - The more we know about the underlying measurement error mechanisms the better the measurement model can be specified
  - The better the measurement model the more effective the adjustment
  - Even if we have an educated guess, it is still worth doing as a sensitivity analysis
- The hardest to implement?
  - To really exploit the full flexibility of Bayesian methods we need to use Bayesian software
  - Stan, WinBUGS

Bayesian Adjustments



SIMEX

- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)

SIMEX

- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)
- The SIMulation-EXtrapolation algorithm
  - Assuming $Y = \alpha + \beta X^* + \epsilon$, and $X^* = X + U$

SIMEX

- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)
- The SIMulation-EXtrapolation algorithm
  - Assuming $Y = \alpha + \beta X^* + \epsilon$, and $X^* = X + U$

1. Generate new variables with increasing levels of measurement error, $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$, with $\lambda_k = (0.5, 1, 1.5, 2)$

SIMEX

- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)
- The SIMulation-EXtrapolation algorithm
  - Assuming $Y = \alpha + \beta X^* + \epsilon$, and $X^* = X + U$

  1. Generate new variables with increasing levels of measurement error, $X^*_k(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$, with $\lambda_k = (0.5, 1, 1.5, 2)$
  2. Re-estimate the outcome model using X^*_k, which produces increasingly biased estimates, $\hat{\beta}^*_k = \beta_1 \frac{S^2_X}{S^2_X + (1+\lambda_k)S^2_U}$

SIMEX

- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)
- The SIMulation-EXtrapolation algorithm
  - Assuming $Y = \alpha + \beta X^* + \epsilon$, and $X^* = X + U$

1. Generate new variables with increasing levels of measurement error, $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$, with $\lambda_k = (0.5, 1, 1.5, 2)$
2. Re-estimate the outcome model using X_k^*, which produces increasingly biased estimates, $\hat{\beta}_k^* = \beta_1 \frac{S_X^2}{S_X^2 + (1+\lambda_k)S_U^2}$
3. Steps 1 and 2 are repeated to obtain $\overline{\hat{\beta}_k^*}$, reduce simulation error

SIMEX

- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)
- The SIMulation-EXtrapolation algorithm
  - Assuming $Y = \alpha + \beta X^* + \epsilon$, and $X^* = X + U$

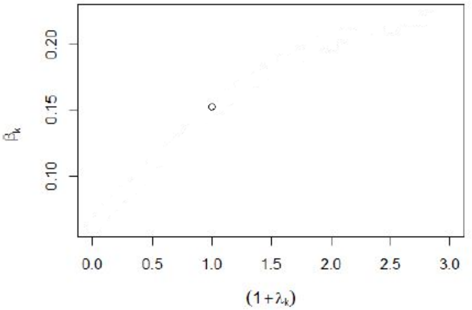
1. Generate new variables with increasing levels of measurement error, $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$, with $\lambda_k = (0.5, 1, 1.5, 2)$
2. Re-estimate the outcome model using X_k^*, which produces increasingly biased estimates, $\hat{\beta}_k^* = \beta_1 \frac{S_X^2}{S_X^2+(1+\lambda_k)S_U^2}$
3. Steps 1 and 2 are repeated to obtain $\overline{\hat{\beta}_k^*}$, reduce simulation error
4. $\overline{\hat{\beta}_k^*}$ and λ_k can now be paired and their relationship estimated

SIMEX

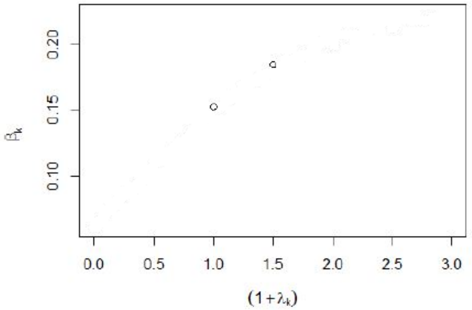
- The simplest approach?
  - Estimates the impact of measurement error through simulation
  - All we need is an understanding of the form and prevalence of the measurement error
  - An R package (simex) with built-in commands to explore the impact in general cases (e.g. additive errors, misclassification)
- The SIMulation-EXtrapolation algorithm
  - Assuming $Y = \alpha + \beta X^* + \epsilon$, and $X^* = X + U$

1. Generate new variables with increasing levels of measurement error, $X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U$, with $\lambda_k = (0.5, 1, 1.5, 2)$
2. Re-estimate the outcome model using X_k^*, which produces increasingly biased estimates, $\hat{\beta}_k^* = \beta_1 \frac{S_X^2}{S_X^2 + (1+\lambda_k) S_U^2}$
3. Steps 1 and 2 are repeated to obtain $\overline{\hat{\beta}_k^*}$, reduce simulation error
4. $\overline{\hat{\beta}_k^*}$ and λ_k can now be paired and their relationship estimated
5. $\hat{\beta}_{SIMEX}$ can now be calculated by extrapolating to $\lambda_k = -1$

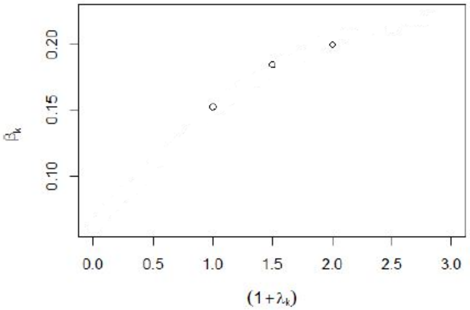
SIMEX



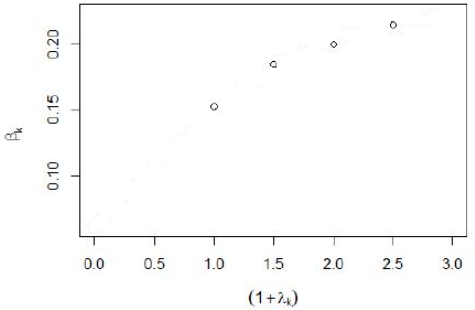
SIMEX



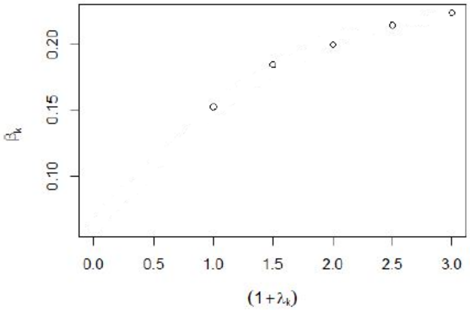
SIMEX



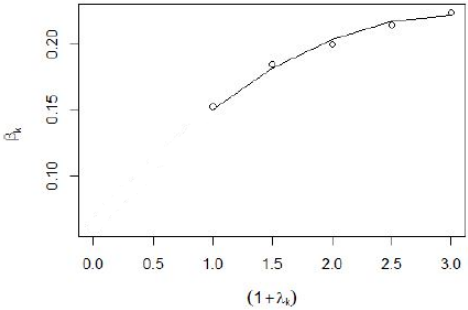
SIMEX



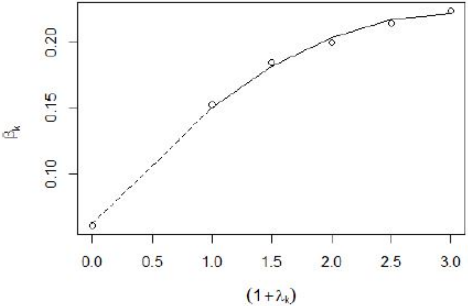
SIMEX



SIMEX



SIMEX



Conclusion

- Many of the variables we use are affected by measurement error
  - Yet, we do very little about it

Conclusion

- Many of the variables we use are affected by measurement error
  - Yet, we do very little about it
- Measurement error can bias not just univariate statistics but associations with other variables perfectly measured
  - Even if the errors are random
  - The specific impact is hard to anticipate beyond simple scenarios

Conclusion

- Many of the variables we use are affected by measurement error
  - Yet, we do very little about it
- Measurement error can bias not just univariate statistics but associations with other variables perfectly measured
  - Even if the errors are random
  - The specific impact is hard to anticipate beyond simple scenarios
- A much bigger problem than we tend to acknowledge
  - Affecting countless studies, potentially quite severely

Conclusion

- Many of the variables we use are affected by measurement error
  - Yet, we do very little about it
- Measurement error can bias not just univariate statistics but associations with other variables perfectly measured
  - Even if the errors are random
  - The specific impact is hard to anticipate beyond simple scenarios
- A much bigger problem than we tend to acknowledge
  - Affecting countless studies, potentially quite severely
- Besides improving data collection processes, we also need to employ adjustments methods
  - Many require additional data, others are quite complex
  - But there are a few methods that are simple enough, and can be used even as a sensitivity analysis tool
  - We should use them every time we suspect measurement error is present in any of our variables

RMeF


2021 Research Methods
e
Festival
Hosted by NCRM and
methods@manchester