Introduction to Measurement Error: Prevalence, Impact and Adjustments

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

Introduction

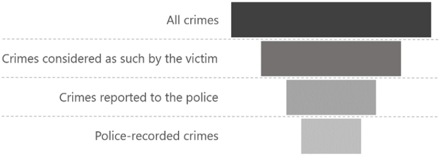
- What is measurement error?
  - Discrepancies between the ‘true’ and the observed value
  - The consequence of a poorly defined construct and/or an imperfect measurement process

Introduction

- What is measurement error?
  - Discrepancies between the ‘true’ and the observed value
  - The consequence of a 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?*)
  - Administrative/official data used as proxies:
    e.g. using earnings to measure poverty, or measuring violent crime from police records

The Recounting Crime project

- Police recorded crime statistics are deeply flawed



  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces

Implications and Solutions

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

    e.g. What is the true prevalence of crime in Hong Kong? Has it grown from last year? Is it higher than in Macau?
  - But also, our causal inferences will be biased

    e.g. Does unemployment affect crime? Does crime affect mental health?

Implications and Solutions

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

    e.g. What is the true prevalence of crime in Hong Kong? Has it grown from last year? Is it higher than in Macau?
  - But also, our causal inferences will be biased

    e.g. Does unemployment affect crime? Does crime affect mental health?
- 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, $U \sim N(0, \sigma_U)$

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

- Examples:
  - Recording inconsistencies due to police officers' recall errors
  - Results from a math test
- Only the variance is affected
  - $\sigma^2_{X^*} = \sigma^2_X + \sigma^2_U$; but the mean is unaffected since $E(U) = 0$

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 $E(U) \neq 0$

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 $E(U) \neq 0$

- Examples:
  - Crime reported to the police
  - Self-reported xenophobia, sexual partners, etc.

Multiplicative Errors

- So far we have assumed that errors are independent
- What if the error is proportional to the true value of the quantity being measured?
  - E.g. memory failures in reporting counts;

    How many alcoholic drinks did you have last week?

Multiplicative Errors

- So far we have assumed that errors are independent
- What if the error is proportional to the true value of the quantity being measured?
  - E.g. memory failures in reporting counts;

    How many alcoholic drinks did you have last week?

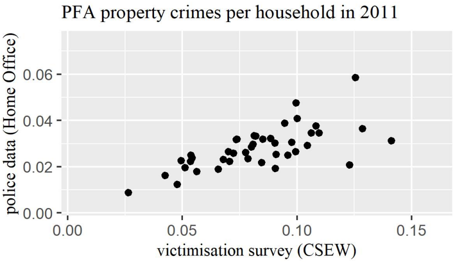
Multiplicative Errors

- So far we have assumed that errors are independent
- What if the error is proportional to the true value of the quantity being measured?
  - E.g. memory failures in reporting counts;
    How many alcoholic drinks did you have last week?
- These can be better specified using a multiplicative rather than an additive model
  - I.e., as $X^* = X \cdot U$, rather than $X^* = X + U$

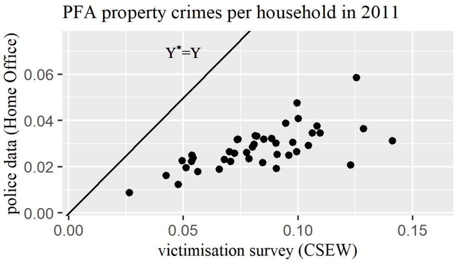
Multiple Error Mechanisms

- Often variables are affected by multiple measurement error mechanisms
- This is how we define measurement error in police data
  - *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

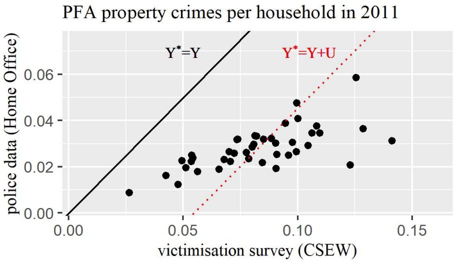
Multiplicative Errors: Crime Rates



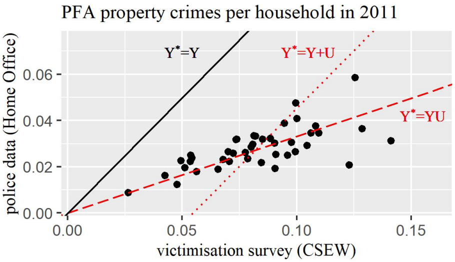
Multiplicative Errors: Crime Rates



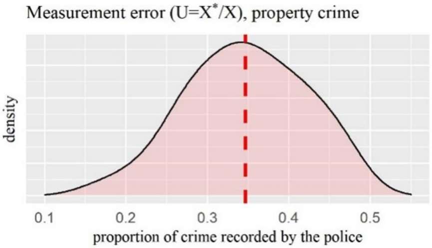
Multiplicative Errors: Crime Rates



Multiplicative Errors: Crime Rates



Multiplicative Errors: Crime Rates



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

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

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

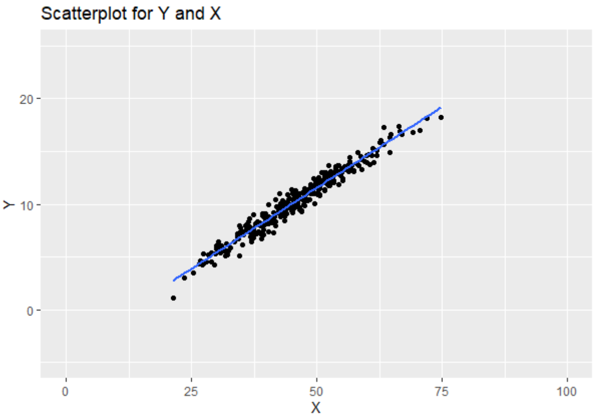
Classical Error on the Response

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

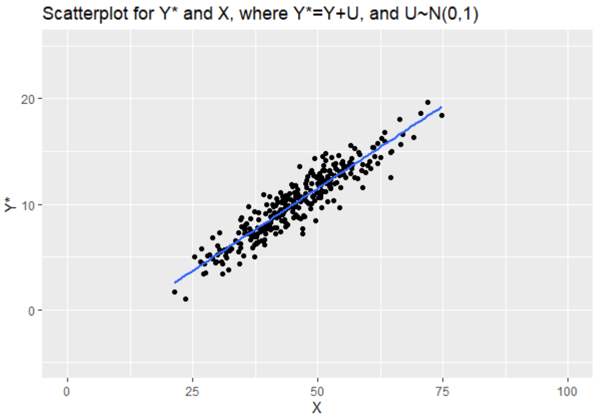
    $Y + U = \alpha + \beta X + \epsilon$

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

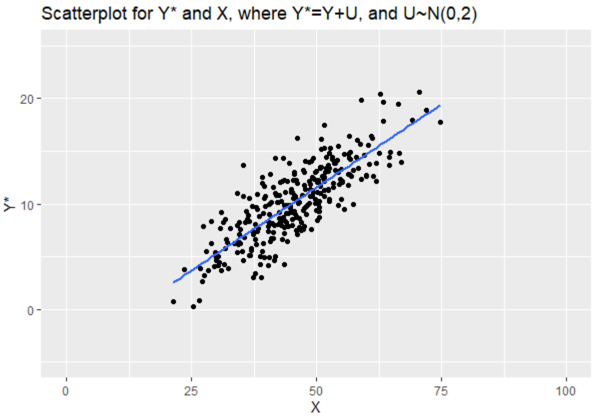
Classical Error on the Response



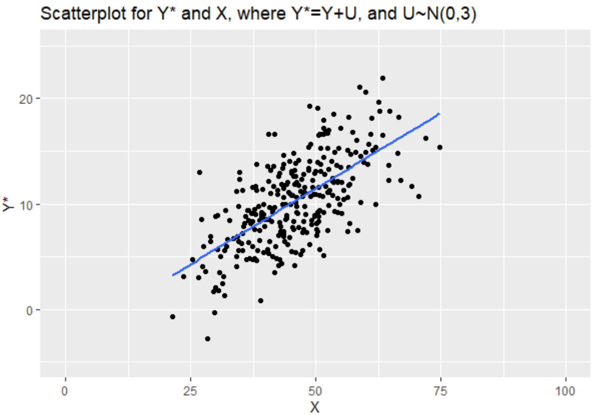
Classical Error on the Response



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...

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

  - 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...

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

  - If instead we have...then..

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

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...

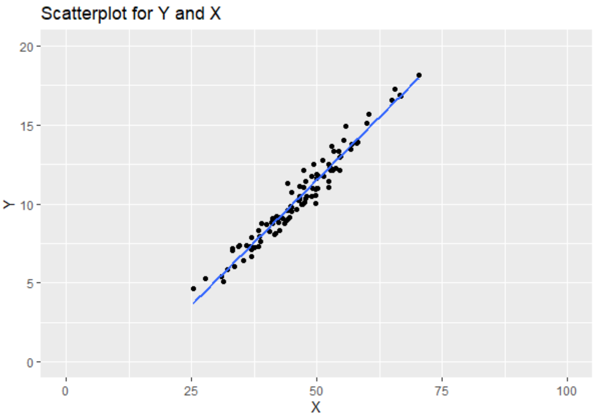
$$\begin{cases} \widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X} \\ \widehat{\beta} = \dfrac{\sigma_{XY}}{\sigma_X^2} \end{cases}$$

  - If instead we have...then..

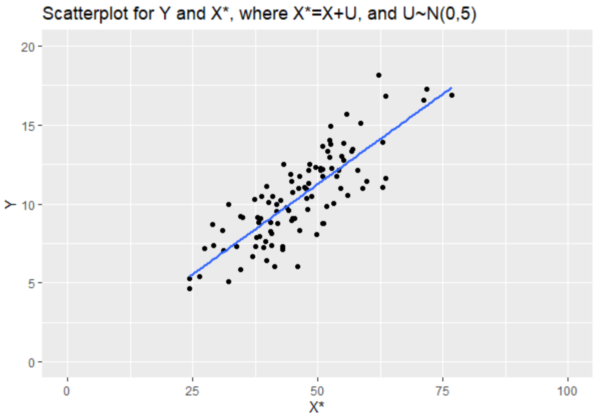
$$\begin{cases} \widehat{\alpha}^* = \bar{Y} - \widehat{\beta}\bar{X}^* = \bar{Y} - \widehat{\beta}\bar{X} = \widehat{\alpha}; & \underline{\text{unbiased constant}} \\ \widehat{\beta}^* = \dfrac{\sigma_{X^*Y}}{\sigma_{X^*}^2} = \dfrac{\sigma_{XY}}{\sigma_X^2 + \sigma_U^2} = \widehat{\beta}\left(\dfrac{\sigma_X^2}{\sigma_X^2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

  - We can see this effect using simulated data

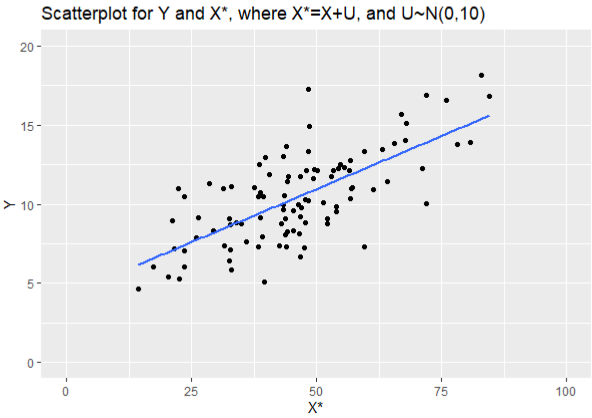
Effect of Random Measurement Error



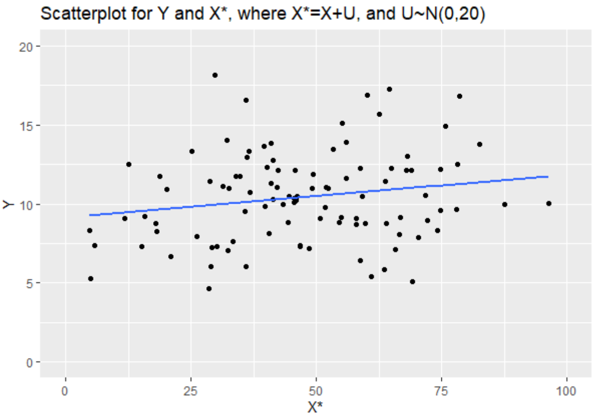
Effect of Random Measurement Error



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
- And we have only considered relatively simple scenarios

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
- And we have only considered relatively simple scenarios
- 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"*

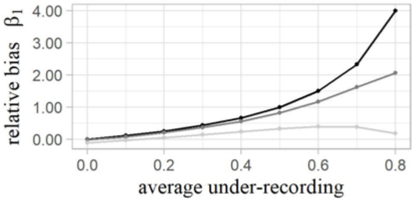
Impact of Errors in Police Data

- What happens when you use police data (subject to multiplicative, systematic and random errors) in linear regression?

Impact of Errors in Police Data

- What happens when you use police data (subject to multiplicative, systematic and random errors) in linear regression?

$$\text{worry} = \alpha + \beta_1 \text{crime}^* + \beta_2 \text{disorder} + \varepsilon$$



Random error: none, moderate (sd=0.04), substantial (sd=0.08)

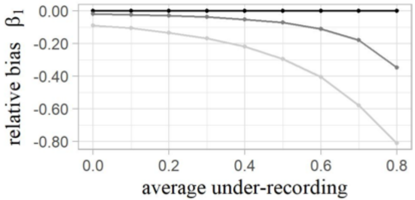
Impact of Errors in Police Data

- What happens when you use police data (subject to multiplicative, systematic and random errors) in linear regression?
- One of our recommendations is to log-transform crime since $log(X^*) = log(X \cdot U) = log(X) + log(U)$

Impact of Errors in Police Data

- What happens when you use police data (subject to multiplicative, systematic and random errors) in linear regression?
- One of our recommendations is to log-transform crime since $log(X^*) = log(X \cdot U) = log(X) + log(U)$

$$\text{worry} = \alpha + \beta_1 \log(\text{crime}^*) + \beta_2 \text{disorder} + \varepsilon$$



Random error: none, moderate (sd=0.04), substantial (sd=0.08)

Adjustment Methods

- We should always aim to improve data collection processes
- When that is not possible/sufficient we should adjust for the impact of measurement error

Adjustment Methods

- We should always aim to improve data collection processes
- When that is not possible/sufficient we should adjust for the impact of measurement error
- We have seen how in some simple settings we can anticipate - and therefore adjust - that impact
- When we can't trace out the impact of measurement error algebraically we need to use other methods

Adjustment 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

- 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

- 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)
  - Simulations (the *RCME* package Pina-Sánchez et al., 2022)

Sensitivity Analysis Using RCME

- We want to estimate the effect of collective efficacy on criminal damage rates, controlling for a few area characteristics
  - $log(Y^*) = \alpha + \beta_1 X + \beta_k Z_k + \epsilon$
  - Population: 982 postcode areas in London
- We estimate it first using police data (naive model), and we re-estimate it using adjusted crime rates (sensitivity analysis)

Sensitivity Analysis Using RCME

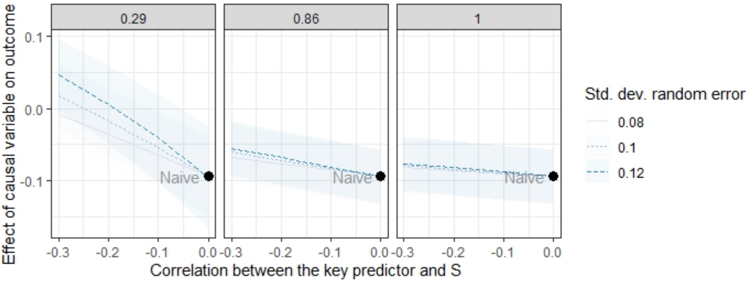
- We want to estimate the effect of collective efficacy on criminal damage rates, controlling for a few area characteristics
  - $log(Y^*) = \alpha + \beta_1 X + \beta_k Z_k + \epsilon$
  - Population: 982 postcode areas in London
- We estimate it first using police data (naive model), and we re-estimate it using adjusted crime rates (sensitivity analysis)
- We simulate different combinations of measurement error to obtain multiple sets of 'adjusted' crime rates
  - We consider values for S (systematic error), R (random error), and D (differential errors)
  - which we derive from the Crime Survey for England and Wales
  - The measurement error term: $U = S + R + D \cdot X$
  - The adjusted crime rates: Y^*/U

The Naive Model

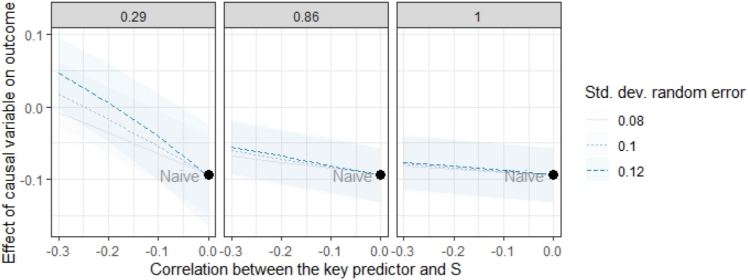
	outcome variable: log(criminal damage)	
	Using police data	Adjusting for measurement error
constant	0.88***	
collective efficacy	**-0.09*****	**?**
% white British	0.11***	
% unemployed	0.14***	
median age	−0.12***	
R^2	0.23	
N	982	

- We re-estimate that model considering different errors
  - $S = (1, 0.86, 0.29)$
  - $R = (0.08, 0.1, 0.12)$
  - $D = (0, -0.1, -0.2, -0.3)$

The *RCME* Adjustment



The *RCME* Adjustment



- The effect size attributed to collective efficacy appears robust, unless...
  - in the presence of low recording rates, variable recording rates across areas, and some minor differential error
  - That is however the most likely scenario according to victimisation surveys

Conclusion

- Many of the variables we use are affected by measurement error

Conclusion

- Many of the variables we use are affected by measurement error
- Measurement error can bias univariate statistics but also 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
- Measurement error can bias univariate statistics but also 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
  - And yet, we do very little about it

Conclusion

- Many of the variables we use are affected by measurement error
- Measurement error can bias univariate statistics but also 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
  - And yet, we do very little about it
- 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 (e.g. simulations) that are simple enough, and can be used even as a sensitivity analysis tool

Thank you!