UNIVERSITY OF LEEDS

SDAI
Secondary Data Analysis Initiative

Crime Surveys User Conference 2020

The impact of measurement error in police crime records

Recounting Crime

Jose Pina-Sánchez (University of Leeds)
Ian Brunton-Smith (University of Surrey)
David Buil-Gil (University of Manchester)
Alexandru Cernat (University of Manchester)

The Problem

- We are all aware that police recorded crime is deeply flawed
  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces

The Problem

- We are all aware that police recorded crime is deeply flawed
  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces
- Yet, we still use police recorded crime rates in our research
  - Key variable to assess the causes and consequences of crime
  - Used by criminologists, economists, geographers, ...

The Problem

- We are all aware that police recorded crime is deeply flawed
  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces
- Yet, we still use police recorded crime rates in our research
  - Key variable to assess the causes and consequences of crime
  - Used by criminologists, economists, geographers, ...
- Multivariate models based on police data will likely be biased
  - These biases can be affecting estimates beyond the effect of crime
  - Their extent will be determined by the type of measurement error and outcome model

The Problem

- We are all aware that police recorded crime is deeply flawed
  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces
- Yet, we still use police recorded crime rates in our research
  - Key variable to assess the causes and consequences of crime
  - Used by criminologists, economists, geographers, ...
- Multivariate models based on police data will likely be biased
  - These biases can be affecting estimates beyond the effect of crime
  - Their extent will be determined by the type of measurement error and outcome model
- In this paper (and project) we seek to tackle this problem
  - Identify the nature and prevalence of measurement error in police data
  - Illustrate the impact it has when used in regression models
  - Suggest methods for its adjustment

Measurement Error in Police Data

- We focus on crime counts/rates
  - We compare police data (data.police.uk) and CSEW estimates of acquisitive crime at the Police Force Area level per year
  - Assuming the latter is a gold standard

Measurement Error in Police Data

- We focus on crime counts/rates
  - We compare police data (data.police.uk) and CSEW estimates of acquisitive crime at the Police Force Area level per year
  - Assuming the latter is a gold standard
- To estimate the prevalence of the measurement error first we need to consider its nature
  - We can anticipate systematic (under-reporting/under-detection) and random (inconsistencies across forces) errors
  - And that these errors are multiplicative (proportional to the true value), not additive

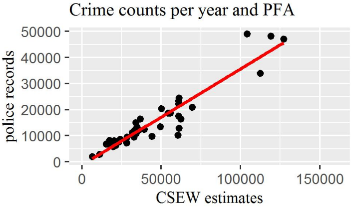
Measurement Error in Police Data

- We focus on crime counts/rates
  - We compare police data (data.police.uk) and CSEW estimates of acquisitive crime at the Police Force Area level per year
  - Assuming the latter is a gold standard
- To estimate the prevalence of the measurement error first we need to consider its nature
  - We can anticipate systematic (under-reporting/under-detection) and random (inconsistencies across forces) errors
  - And that these errors are multiplicative (proportional to the true value), not additive

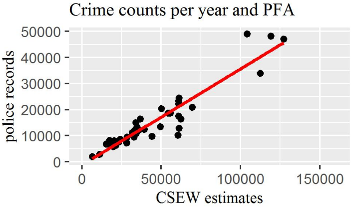
    additive error model: $X^* = X + U$, where $U \sim N(< 0, \sigma)$

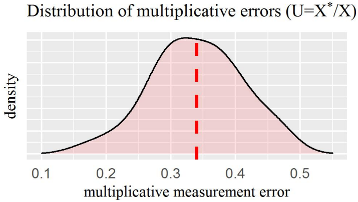
    multiplicative error model: $X^* = X \cdot U$, where $U \sim N(\in (0,1), \sigma)$

Measurement Error in Police Data

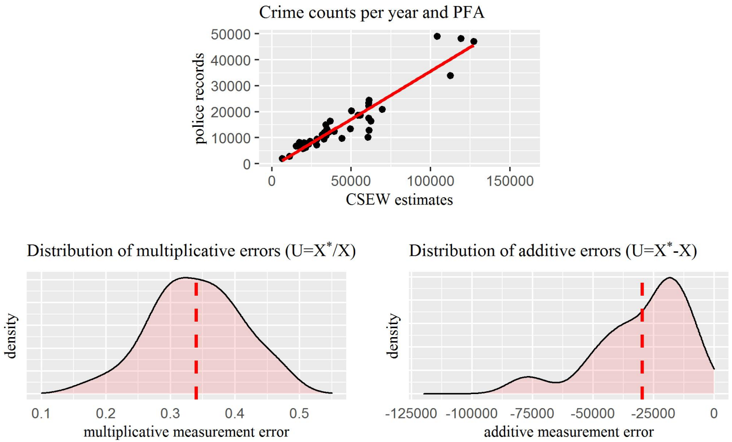


Measurement Error in Police Data





Measurement Error in Police Data



Impact of Multiplicative Errors

- The nature of the measurement error (not only its prevalence) defines its impact
- Let's assume systematic and random measurement error on the explanatory variable of a simple linear model

$$Y = \alpha + \beta X^* + \epsilon$$

Impact of Multiplicative Errors

- The nature of the measurement error (not only its prevalence) defines its impact
- Let's assume systematic and random measurement error on the explanatory variable of a simple linear model

$$Y = \alpha + \beta X^* + \epsilon$$

- Using OLS we can estimate α and β solving the following system of equations

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}^*$$

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?

  1. Random noise in X^* doesn't affect cov but increases var $\rightarrow$ attenuates the slope proportional to the reliability ratio

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?
  1. Random noise in X^* doesn't affect cov but increases var $\rightarrow$ attenuates the slope proportional to the reliability ratio
  2. Neither cov nor var are affected by a change of origin $\rightarrow$ systematic additive error won't bias the slope

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?
  1. Random noise in X^* doesn't affect *cov* but increases *var* → attenuates the slope proportional to the reliability ratio
  2. Neither *cov* nor *var* are affected by a change of origin → systematic additive error won't bias the slope
- What happens if the errors are multiplicative?

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?
  1. Random noise in X^* doesn't affect *cov* but increases *var* → attenuates the slope proportional to the reliability ratio
  2. Neither *cov* nor *var* are affected by a change of origin → systematic additive error won't bias the slope
- What happens if the errors are multiplicative?
  1. Random noise in X^* doesn't affect *cov*, but increases *var* → attenuates the slope proportional to the reliability ratio

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?
  1. Random noise in X^* doesn't affect *cov* but increases *var* $\rightarrow$ attenuates the slope proportional to the reliability ratio
  2. Neither *cov* nor *var* are affected by a change of origin $\rightarrow$ systematic additive error won't bias the slope
- What happens if the errors are multiplicative?
  1. Random noise in X^* doesn't affect *cov*, but increases *var* $\rightarrow$ attenuates the slope proportional to the reliability ratio
  2. *cov* and especially *var* are affected by a change of scale $\rightarrow$ systematic multiplicative error will bias the slope; downwards if the errors are +, upwards if -; proportional to the change of scale

Impact on the Slope

- Let's focus on the slope since this is often what we are after

$$\hat{\beta} = \frac{cov(X^*Y)}{var(X^*)}$$

- What happens if the errors are additive?
  1. Random noise in X^* doesn't affect *cov* but increases *var* $\rightarrow$ attenuates the slope proportional to the reliability ratio
  2. Neither *cov* nor *var* are affected by a change of origin $\rightarrow$ systematic additive error won't bias the slope
- What happens if the errors are multiplicative?
  1. Random noise in X^* doesn't affect *cov*, but increases *var* $\rightarrow$ attenuates the slope proportional to the reliability ratio
  2. *cov* and especially *var* are affected by a change of scale $\rightarrow$ systematic multiplicative error will bias the slope; downwards if the errors are +, upwards if -; proportional to the change of scale
- Under-reporting will bias the slope upwards, while inconsistencies across PFAs will push it downwards

Adjustments

- If we can estimate the validity and reliability of police data
  - We can anticipate and adjust the impact of measurement error

Adjustments

- If we can estimate the validity and reliability of police data
  - We can anticipate and adjust the impact of measurement error
- However, things become more complicated when we move away from simple linear regression
  - The errors in crime rates could affect bias the slopes of other variables included in the model
  - Harder to trace out if using non-linear models
  - And we have not even consider how it affects measures of uncertainty too

Adjustments

- If we can estimate the validity and reliability of police data
  - We can anticipate and adjust the impact of measurement error
- However, things become more complicated when we move away from simple linear regression
  - The errors in crime rates could affect bias the slopes of other variables included in the model
  - Harder to trace out if using non-linear models
  - And we have not even consider how it affects measures of uncertainty too
- We recommend using Bayesian adjustments
  - We need to specify our outcome model of interest
  - Together with a measurement model for the variable affected by measurement error
  - Based on the estimated validity and reliability of that variable

Example of a Bayesian Adjustment

- We model the effect of acquisitive crime (X_1) and population density (X_2) on the % of white population (Y)
  - We do that in three steps

Example of a Bayesian Adjustment

- We model the effect of acquisitive crime (X_1) and population density (X_2) on the % of white population (Y)
  - We do that in three steps
- The (assumed) true model, using CSEW data
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Example of a Bayesian Adjustment

- We model the effect of acquisitive crime (X_1) and population density (X_2) on the % of white population (Y)
  - We do that in three steps
- The (assumed) true model, using CSEW data
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- The naive model, using police data
  - $Y = \alpha + \beta_1 X_1^* + \beta_2 X_2 + \epsilon$

Example of a Bayesian Adjustment

- We model the effect of acquisitive crime (X_1) and population density (X_2) on the % of white population (Y)
  - We do that in three steps
- The (assumed) true model, using CSEW data
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
- The naive model, using police data
  - $Y = \alpha + \beta_1 X_1^* + \beta_2 X_2 + \epsilon$
- The Bayesian adjustment, using police data and what we know about the validity and reliability of the measurement error
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
    $X_1 = X_1^* \cdot U$, where $U \sim N(1/0.29, 1/0.06)$

Example Using CSEW and Police Data

	outcome variable: % white in the area		
	true model	naive model	adjusted model
constant	0.977		
	(0.013)		
acquisitive crime	−0.073		
	(0.029)		
population density	−0.003		
	(0.002)		
observations	40	40	40

Example Using CSEW and Police Data

	outcome variable: % white in the area		
	true model	naive model	adjusted model
constant	0.977	0.976	
	(0.013)	(0.012)	
acquisitive crime	−0.073	−0.276	
	(0.029)	(0.095)	
population density	−0.003	−0.003	
	(0.002)	(0.002)	
observations	40	40	40

Example Using CSEW and Police Data

	outcome variable: % white in the area		
	true model	naive model	adjusted model
constant	0.977	0.976	0.976
	(0.013)	(0.012)	(0.013)
acquisitive crime	−0.073	−0.276	−0.081
	(0.029)	(0.095)	(0.030)
population density	−0.003	−0.003	−0.002
	(0.002)	(0.002)	(0.002)
observations	40	40	40

Discussion

- The type of measurement error observed in crime rates can be defined as
  - Multiplicative, with a strong negative systematic component, normally distributed across police forces

Discussion

- The type of measurement error observed in crime rates can be defined as
  - Multiplicative, with a strong negative systematic component, normally distributed across police forces
- This type of errors will lead to strong biases when used in regression models
  - If used as an explanatory variable on a linear model it will normally lead to an upward bias for the effect of crime
  - We have shown the impact in one simple type of model
  - The validity of practically any study relying on such data is under question

Discussion

- The type of measurement error observed in crime rates can be defined as
  - Multiplicative, with a strong negative systematic component, normally distributed across police forces
- This type of errors will lead to strong biases when used in regression models
  - If used as an explanatory variable on a linear model it will normally lead to an upward bias for the effect of crime
  - We have shown the impact in one simple type of model
  - The validity of practically any study relying on such data is under question
- We recommend employing Bayesian adjustments
  - Very flexible, can be used with any kind of outcome model, and form of measurement error
  - Can be used as sensitivity tools when all we have is an educated guess of the validity and reliability of police data

Discussion

- The type of measurement error observed in crime rates can be defined as
  - Multiplicative, with a strong negative systematic component, normally distributed across police forces
- This type of errors will lead to strong biases when used in regression models
  - If used as an explanatory variable on a linear model it will normally lead to an upward bias for the effect of crime
  - We have shown the impact in one simple type of model
  - The validity of practically any study relying on such data is under question
- We recommend employing Bayesian adjustments
  - Very flexible, can be used with any kind of outcome model, and form of measurement error
  - Can be used as sensitivity tools when all we have is an educated guess of the validity and reliability of police data
- Next steps
  - Small area estimation, synthetic data, multi-trait multi-method models, SIMEX
  - `http://recountingcrime.com`