Royal Statistical Society 2021 Conference

The Impact of Measurement Error in Regression Models Using Police Recorded Crime Rates

Recounting Crime

Background

- Police recorded crime data is deeply flawed
  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces
  - Under-recording associated with key variables of interest
- We are all aware of the problem, yet, we still use it everywhere

Background

- Police recorded crime data is deeply flawed
  - Under-reporting/under-detection of crime
  - Recording inconsistencies across forces
  - Under-recording associated with key variables of interest
- We are all aware of the problem, yet, we still use it everywhere
- Common in multivariate models exploring the causes and consequences of crime
  - The effect of inequality, unemployment, racial segregation, police numbers, police practices,... on crime
  - Or the effect of crime on, perceptions of security, outdoors exercise, insurance purchases, gun ownership, ...
  - Used by criminologists, but also: economists, geographers, demographers, sociologists, epidemiologists, ...

Background

- We normally include a few caveats and move on
  - I do, guilty as charged

Background

- We normally include a few caveats and move on
  - I do, guilty as charged
- Now, can we really move on?
  - What are the specific implications of using this data?
  - How biased are our findings?
  - Should we keep publishing studies using police data?
  - Or are we 'polluting' the evidence base?

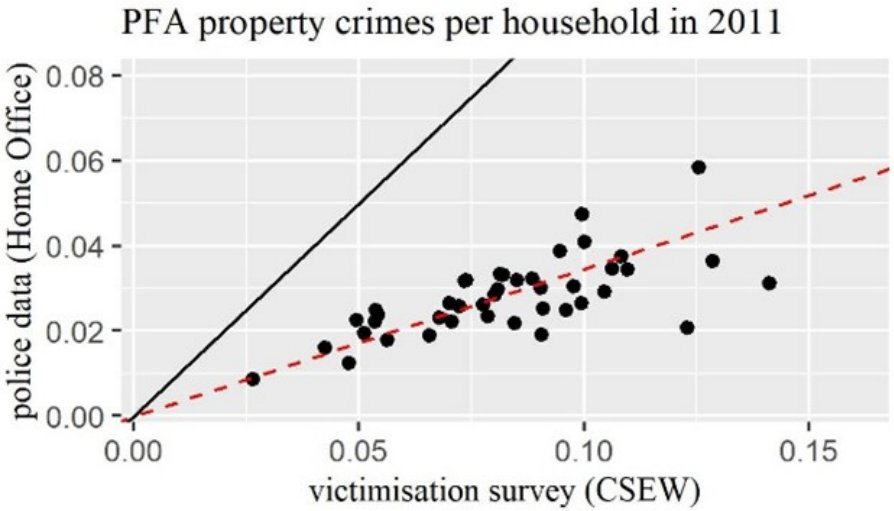
Background

- We normally include a few caveats and move on
  - I do, guilty as charged
- Now, can we really move on?
  - What are the specific implications of using this data?
  - How biased are our findings?
  - Should we keep publishing studies using police data?
  - Or are we 'polluting' the evidence base?
- That's what we seek to explore in *Recounting Crime*
  - Understand the impact of measurement error in police data
  - So we can re-examine findings from the literature more robustly
  - And as much as possible adjust for this problem in the future

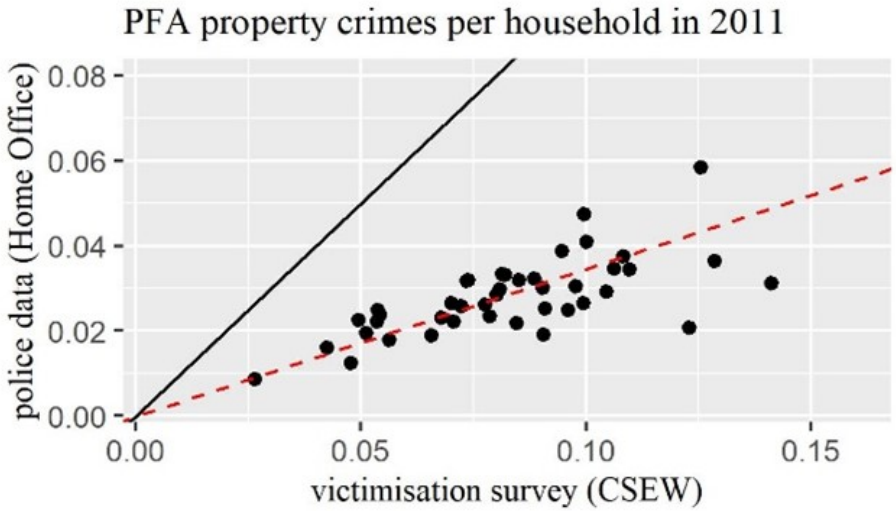
Prevalence of Measurement Error

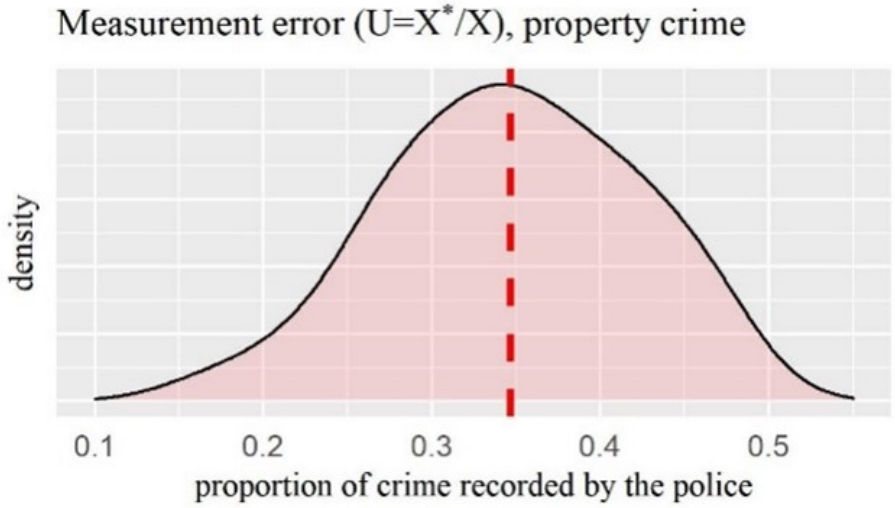
- We started by exploring measurement error in police recorded crime rates across areas
  - We anticipate the errors will be both systematic (under-reporting) and random (inconsistencies across forces)
  - And the functional form will be multiplicative (errors proportional to the true value)
  - $X^* = X \cdot U \; ; \quad U \sim N(\bar{U}, \sigma) \; ; \quad \bar{U} \in (0, 1)$
- To test the above we compared police recorded crime rates against...
  - Crime survey estimates of property crime, across Police Force Areas in England
  - Centre for Disease Control registered homicides, across States in the US
  - Assuming victimisation surveys and vital registers can be considered gold standards

Property Crime Rates (England)

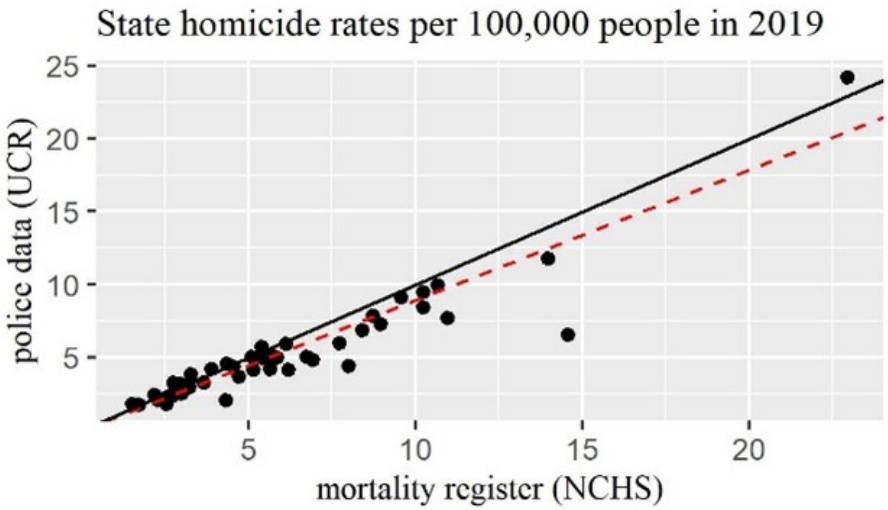


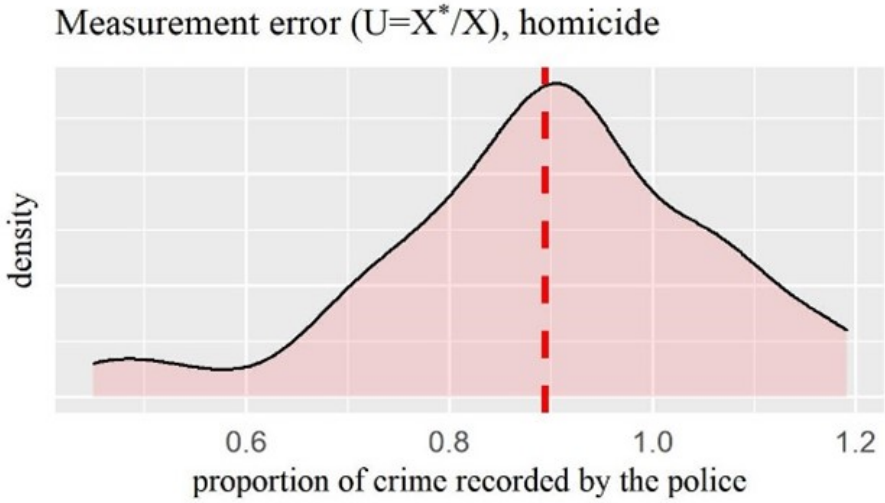
Property Crime Rates (England)





Homicide Rates (US)





Impact in Regression Models

- What would then be the impact of using police data in regression models?
- We tried to explore this formally (using algebra) and empirically (using simulations)

Impact in Regression Models

- What would then be the impact of using police data in regression models?
- We tried to explore this formally (using algebra) and empirically (using simulations)
- We seek to generalise to all settings where police crime rates are used in linear models
- To do so we assume the measurement error is...
  - multiplicative, systematic (negative), normally distributed, non-differential

Crime as the Response Variable

- Let's consider a linear model exploring the causes of crime
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Crime as the Response Variable

- Let's consider a linear model exploring the causes of crime
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - where, $Y^* = Y \cdot U$, and $U \sim N((0,1), \sigma)$

Crime as the Response Variable

- Let's consider a linear model exploring the causes of crime
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - where, $Y^* = Y \cdot U$, and $U \sim N((0,1), \sigma)$
  - then, $Y = \dfrac{\alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon}{U}$
  - regression coefficients biased upwards proportionally to the under-recording rate

Crime as the Response Variable

- Let's consider a linear model exploring the causes of crime
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - where, $Y^* = Y \cdot U$, and $U \sim N((0,1), \sigma)$
  - then, $Y = \dfrac{\alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon}{U}$
  - regression coefficients biased upwards proportionally to the under-recording rate
- The implication is effect sizes reported in the literature are overdimensioned
  - but not all studies are affected by this problem

Crime as the Response Variable

- If crime rates are log-transformed
  - $log(Y^*) = log(Y \cdot U) = log(Y) + log(U)$
  - the multiplicative error model turns additive

Crime as the Response Variable

- If crime rates are log-transformed
  - $log(Y^*) = log(Y \cdot U) = log(Y) + log(U)$
  - the multiplicative error model turns additive
- What is the effect of systematic *additive* error then?
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - where, $Y^* = Y + U$, and $U \sim N(< 0, \sigma)$

Crime as the Response Variable

- If crime rates are log-transformed
  - $log(Y^*) = log(Y \cdot U) = log(Y) + log(U)$
  - the multiplicative error model turns additive
- What is the effect of systematic *additive* error then?
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
  - where, $Y^* = Y + U$, and $U \sim N(< 0, \sigma)$
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon - U$
  - the intercept is biased (upwards), plus we lose some precision
- The implications are possibly Type II errors

Crime as an Explanatory Variable

- Let's consider a simple linear model, where the explanatory variable is affected by measurement error

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

Crime as an Explanatory Variable

- Let's consider a simple linear model, where the explanatory variable is affected by measurement error

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

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

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

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

Crime as an Explanatory Variable

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

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

Crime as an Explanatory Variable

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

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

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

Crime as an Explanatory Variable

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

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

1. Random noise in X^* doesn't affect cov, but increases $var \rightarrow$ attenuates the slope
2. Systematic multiplicative error will lead to a change of scale, which affects the cov and especially $var \rightarrow$ the slope will be biased; augmented if the errors are negative

Crime as an Explanatory Variable

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

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

1. Random noise in X^* doesn't affect cov, but increases $var \rightarrow$ attenuates the slope
2. Systematic multiplicative error will lead to a change of scale, which affects the cov and especially $var \rightarrow$ the slope will be biased; augmented if the errors are negative

- Under-reporting will bias the slope upwards, while inconsistencies across PFAs will push it downwards

Impact in more Complex Settings

- Things get harder when we consider other outcome models
  - When we use non-linear models
  - When we introduce more explanatory variables
  - Or when we consider 'causal models': i.e. fixed effects, instrumental variables, etc.

Impact in more Complex Settings

- Things get harder when we consider other outcome models
  - When we use non-linear models
  - When we introduce more explanatory variables
  - Or when we consider 'causal models': i.e. fixed effects, instrumental variables, etc.
- And so far we have assumed that the measurement error is not associated to any other variable of interest
  - i.e. non-differential errors, $cov(\epsilon, U) = 0$
  - But we know this is not true
  - Well established that under-reporting is associated with deprivation, ethnicity and many other factors
  - What will be the impact of such 'non-differential' errors?

Impact in more Complex Settings

- Things get harder when we consider other outcome models
  - When we use non-linear models
  - When we introduce more explanatory variables
  - Or when we consider 'causal models': i.e. fixed effects, instrumental variables, etc.
- And so far we have assumed that the measurement error is not associated to any other variable of interest
  - i.e. non-differential errors, $cov(\epsilon, U) = 0$
  - But we know this is not true
  - Well established that under-reporting is associated with deprivation, ethnicity and many other factors
  - What will be the impact of such 'non-differential' errors?
- To explore this question we have built a synthetic population
  - Reflecting the victims of crime that we would expect to see according to the Crime Survey

Discussion

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

Discussion

- The type of measurement error observed in police recorded crime rates can be defined as ...
  - multiplicative, with a strong negative systematic component, normally distributed across areas
- Those type of errors can lead to strong biases when present in regression models
  - The specific impact varies massively across different settings
  - Quite problematic if crime rates introduced in their original scale
  - The validity of much of the literature is under question

Discussion

- The type of measurement error observed in police recorded crime rates can be defined as ...
  - multiplicative, with a strong negative systematic component, normally distributed across areas
- Those type of errors can lead to strong biases when present in regression models
  - The specific impact varies massively across different settings
  - Quite problematic if crime rates introduced in their original scale
  - The validity of much of the literature is under question
- When using police data we need to check the potential impact of measurement error
  - We should always use logarithmic transformations
  - We should complement our results with sensitivity analysis
  - All we need is an educated guess of: i) the proportion of under-recording, ii) how much that can vary by area, and iii) the relationship between the variables of interest and the under-recording
  - Which can be inferred using our synthetic data (to be published)

Thank You

- If you want to know more...
  - Preprint: `osf.io/preprints/socarxiv/ydf4b/`
  - Project's website: `http://recountingcrime.com`