Security Lancaster Seminars

How Does Measurement Error in Police Data Affect Research on the Causes and Consequences of Crime?

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

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 police data, widely
  - Used in multivariate models to explore the causes and consequences of crime
  - By criminologists, but also economists, geographers, demographers, sociologists, epidemiologists, ...
  - Exploring 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, ...

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 simply polluting the evidence body?

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 simply polluting the evidence body?
- 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

Measurement Error in Police Data

- Two applications focused on cross-sectional crime counts/rates
- To estimate the impact of measurement error first we need to estimate its prevalence
- And to estimate its prevalence we must first consider its form

Measurement Error in Police Data

- Two applications focused on cross-sectional crime counts/rates
- To estimate the impact of measurement error first we need to estimate its prevalence
- And to estimate its prevalence we must first consider its form
  - 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)
  - $X^* = X \cdot U$

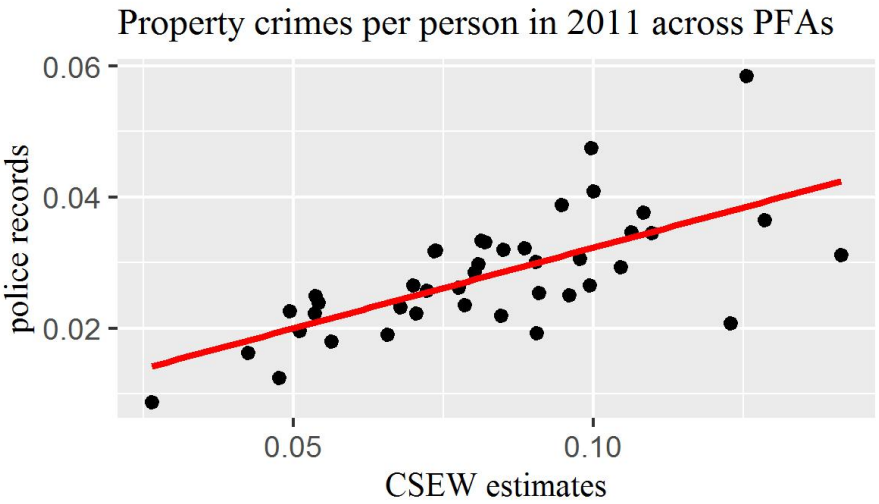
    $U \sim N((0,1), \sigma)$

Measurement Error in Police Data

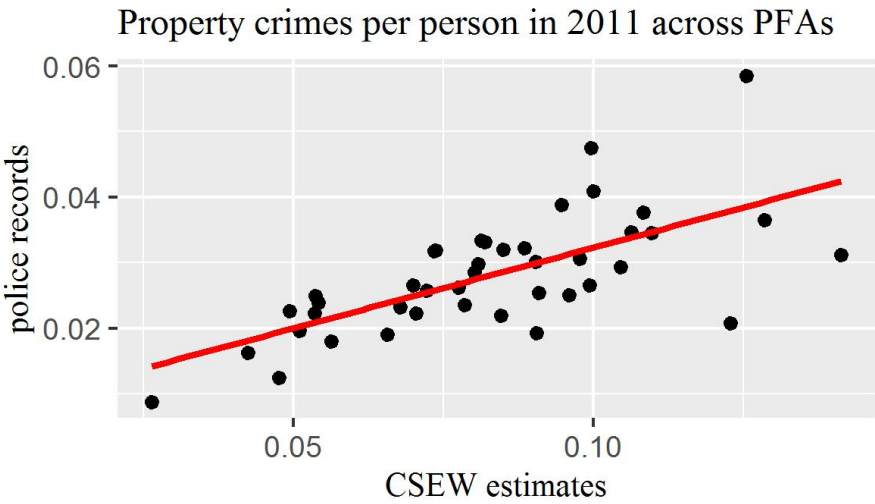
- Two applications focused on cross-sectional crime counts/rates
- To estimate the impact of measurement error first we need to estimate its prevalence
- And to estimate its prevalence we must first consider its form
  - We can anticipate systematic (under-reporting/under-detection) and random (inconsistencies across forces) errors
  - And that these errors are multiplicative (proportional to the true value)
  - $X^* = X \cdot U$

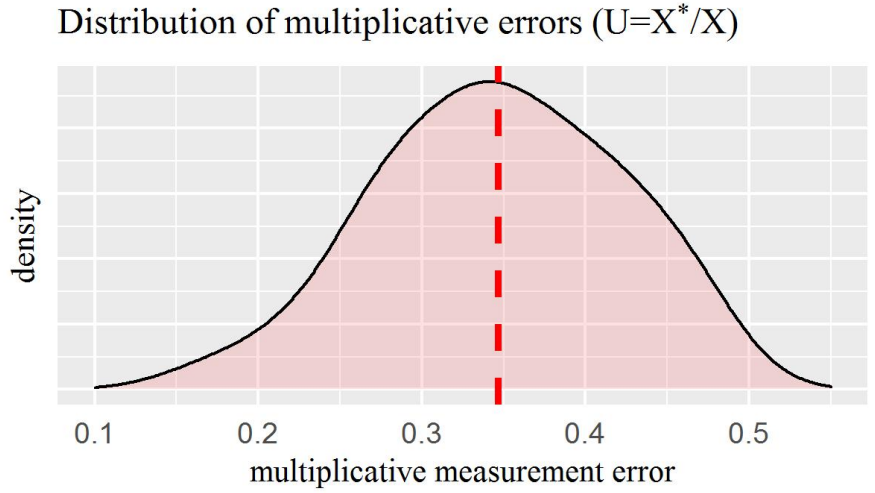
    $U \sim N((0,1), \sigma)$
- To test this we compare police data (data.police.uk) and CSEW estimates
  - Assuming the latter is a gold standard
  - We look at property crime across Police Force Areas

Measurement Error in Police Data



Measurement Error in Police Data





Impact of Measurement Error

- The second step to understanding the impact of measurement error is to consider how will the affected variable be used
- Let's consider a linear model exploring the causes of crime
  - $Y^* = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Impact of Measurement Error

- The second step to understanding the impact of measurement error is to consider how will the affected variable be used
- 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 $U \sim N((0,1), \sigma)$

Impact of Measurement Error

- The second step to understanding the impact of measurement error is to consider how will the affected variable be used
- 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

Impact of Measurement Error

- The second step to understanding the impact of measurement error is to consider how will the affected variable be used
- 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
  - i.e. overdimensioned effect sizes in lots of studies, but not all...

Impact of Under-Recording

- There is an easy fix, log-transform crime rates
  - $log(Y^*) = log(Y \cdot U) = log(Y) + log(U)$
  - Turns the multiplicative into an additive error model

Impact of Under-Recording

- There is an easy fix, log-transform crime rates
  - $log(Y^*) = log(Y \cdot U) = log(Y) + log(U)$
  - Turns the multiplicative into an additive error model
- 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 $U \sim N(< 0, \sigma)$

Impact of Under-Recording

- There is an easy fix, log-transform crime rates
  - $log(Y^*) = log(Y \cdot U) = log(Y) + log(U)$
  - Turns the multiplicative into an additive error model
- 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 $U \sim N(< 0, \sigma)$
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon - U$
  - Only the intercept is biased (upwards)

Impact in more Complex Settings

- Things get harder when we consider other outcome models
  - When crime rates are used as a regressor the systematic and random components of the measurement error will create different biases
  - Taking logs isn't a silver bullet any more
  - Non-linear models add even more complexity

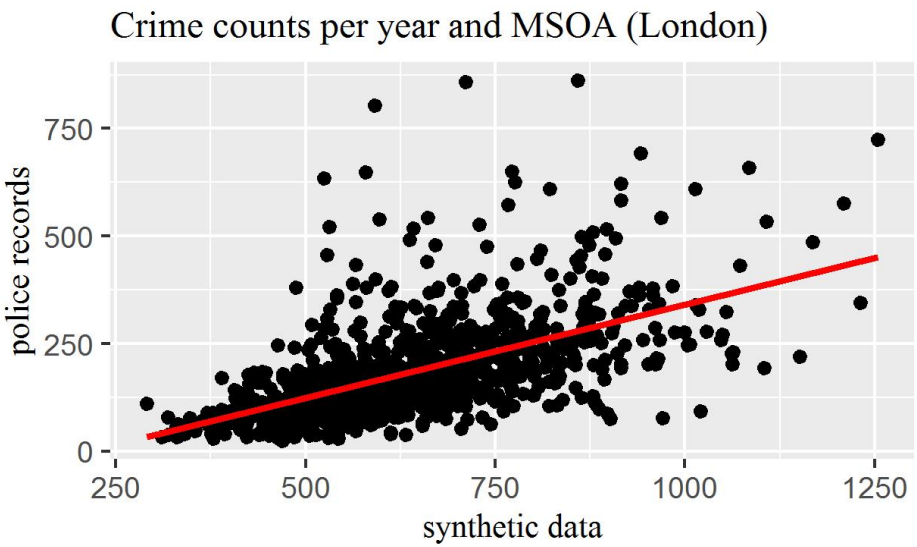
Impact in more Complex Settings

- Things get harder when we consider other outcome models
  - When crime rates are used as a regressor the systematic and random components of the measurement error will create different biases
  - Taking logs isn't a silver bullet any more
  - Non-linear models add even more complexity
- 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?

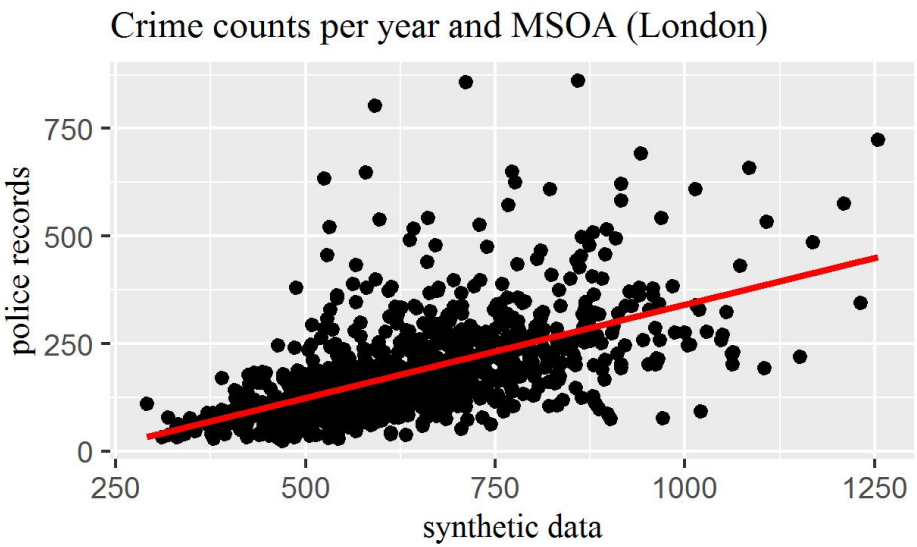
Synthetic Crime Data

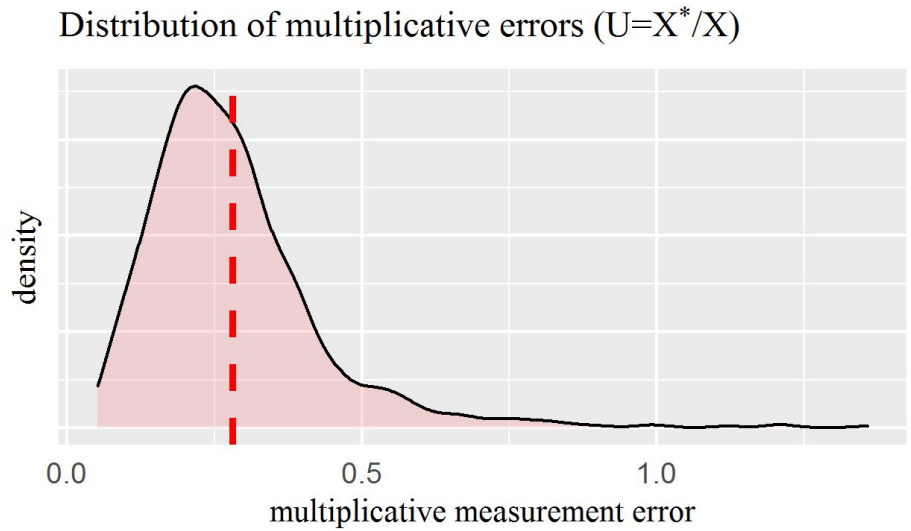
- We generate synthetic data reflecting the distribution of crime rates across England
  - Using victimisation reports from the CSEW and the Census
  - We estimate the distribution of crimes that could be derived if the CSEW sample covered the whole population of England
- We focus on violent crime across London's MSOAs
  - Compare synthetic crime rates to those recoded by the police

Measurement Error in Police Data



Measurement Error in Police Data





Non-Differential Errors

- We find the expected systematic under-recording
  - But as before this is not consistent across areas

Non-Differential Errors

- We find the expected systematic under-recording
  - But as before this is not consistent across areas
- We find all kinds of area-level characteristics associated with crime recording rates

Non-Differential Errors

- We find the expected systematic under-recording
  - But as before this is not consistent across areas
- We find all kinds of area-level characteristics associated with crime recording rates

median income	% unemployment	% no qualification
-0.36	0.39	0.15

% white residents	residential density	collective efficacy
-0.38	0.22	-0.40

Impact of Non-Differential Errors

- We know the synthetic data is not a gold standard
  - Non-response bias, social desirability, interviewer effects, ...
  - Discrepancies between the police and victims in what constitutes a crime
  - Reflecting a specific point in time (2011) and place (England)
- Still, can be useful as a sensitivity analysis tool
  - We suggest using some of these correlations to impute new crime counts with which re-estimate our models of interest
  - And in so doing illustrate the potential impact of non-differential errors in crime rates

Example Sensitivity Analysis

- Let's imagine we use police data to estimate the following
  - $Violence = \alpha + \beta_1 Unemployment + \beta_2 CollectiveEff + \epsilon$
- To proceed with the sensitivity analysis we can
  - Impute violent crime based on the association between unemployment and the measurement error
  - Replicate the model replacing police recorded violent crimes by our imputations
  - We could do this for a range of associations since we do not know the exact true value
  - We could also do this based on more than one factor to increase the accuracy of the imputed values

Example Sensitivity Analysis

	outcome variable: violent crimes	
	Using police data	Using imputed data
constant	**6.91**	
	(0.29)	
unemployment	**8.10**	
	(0.71)	
collective efficacy	**-3.63**	
	(0.37)	
R^2	0.38	
N	980	

Example Sensitivity Analysis

	outcome variable: violent crimes	
	Using police data	Using imputed data
constant	**6.91**	**8.96**
	(0.29)	(0.29)
unemployment	**8.10**	−0.80
	(0.71)	(0.71)
collective efficacy	**-3.63**	**-3.63**
	(0.37)	(0.37)
R^2	0.38	0.12
N	980	980

Discussion

- The type of measurement error observed in crime rates can be defined as
  - Multiplicative, with a strong negative systematic component, normally distributed across areas
  - Associated to variables used to explore the causes and consequences of crime

Discussion

- The type of measurement error observed in crime rates can be defined as
  - Multiplicative, with a strong negative systematic component, normally distributed across areas
  - Associated to variables used to explore the causes and consequences of crime
- These type of errors can lead to strong biases when used in regression models
  - The validity of much of the literature 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 areas
  - Associated to variables used to explore the causes and consequences of crime
- These type of errors can lead to strong biases when used in regression models
  - The validity of much of the literature relying on such data is under question
- When using police data we need to check the potential impact of measurement error
  - Should always use logarithmic transformations to turn multiplicative into additive errors
  - Sensitivity analysis can be really useful to explore the effect of non-differential errors
  - All we need is an educated guess of the relationship between the variables of interest and 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`