Adjusting for systematic and differential errors using simulations

Jose Pina-Sánchez, Ian Brunton-Smith, David Buil-Gil, Alexandru Cernat, and Albert Varela

09 January, 2024

Introduction

This is the first practical exercise from day 2 of the short course 'Adjustment Methods for Data Quality Problems: Missing Data, Measurement Error, and Misclassification.' You can download the associated lecture slides and data here: https://github.com/jmpinasanchez/measurement

In this workshop we are going to see how we can adjust for systematic and differential forms of measurement error using simulations. This is the least sophisticated way of adjusting for measurement error, but as we will see it is quite flexible in that it can be used to represent a wide range of measurement error processes. This approach involves modifying the error-prone variable in order to strip away the measurement error process at play, so we can retrieve the values of the unobserved true variable.

We will practice how to do this through a series of examples exploring some of the forms of measurement error present in police recorded crime data (Pina-Sánchez et al., 2023a). As shown in Figure 1, police recorded crime is bound to systematically underrepresent the true extent of crime since not all crime is reported to - or detected by - the police. We will see how these systematic errors can be easily approximated and accounted for in our adjustments using estimates from crime surveys. We will then proceed to explore more complex differential errors, which could take place when the key variable of interest in our analysis (i.e. the variable explored as either a cause or effect of crime) is also affecting police recording.

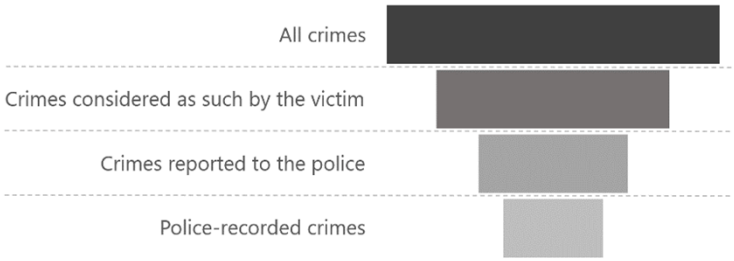


Figure 1: The dark figure of crime

To practice the adjustment for systematic and differential errors we will use criminal damage rates (per 1000 residents) recorded in 2011, across Middle layer Super Output Areas (MSOAs) in Greater London (N = 982). These crime rates are derived from data.police.uk and the 2011 Census. Using further data from the Census and from the Metropolitan Police Service Public Attitudes Survey 2011–2013, we will explore the robustness of the effect ethnic homogeneity and collective efficacy on crime. We will assess the robustness of ethnic homogeneity together, step by step, then you will be asked to repeat that process on your own for the case of unemployment. We will also see how these types of adjustments for non-random errors such as those seen in police data can be facilitated using the `rcme` package (Pina-Sánchez, 2023b). We conclude the practical demonstrating the key limitation of adjustments based on simulations of measurement error mechanisms, namely, that we cannot retrieve the true value of variables subject to random errors.

Exercise 1. Adjusting for systematic errors in police data

Let's start by importing the data, and taking a quick look at it.

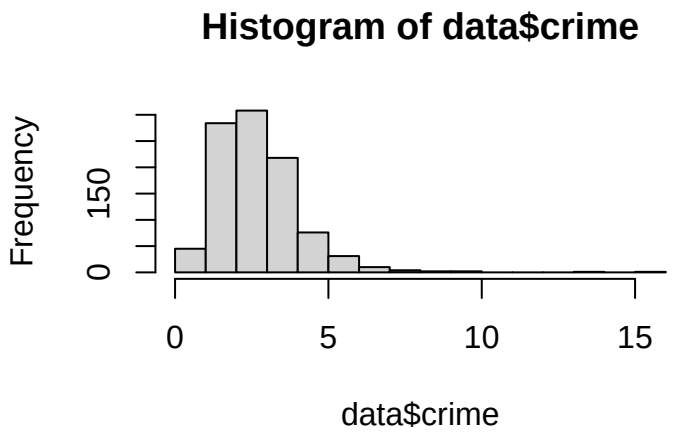
```
#Importing the data.
data = read.csv("crime_London.csv")
#Remember to use the address of the folder where you saved the dataset.
head(data)
```

```
##   crime   age collective unemployment  white
## 1  3.83 -0.27     -0.627         1.56  0.233
## 2  2.38  0.20     -0.764         0.61 -0.304
## 3  0.49  0.91      1.124         0.68  1.058
## 4  2.56 -0.27     -0.157         0.82  0.300
## 5  4.40 -0.51     -0.043         1.49  0.083
## 6  3.17 -0.74     -1.758         2.02  0.346
```

```
summary(data)
```

```
##      crime            age          collective     unemployment      white
##  Min.   : 0.4   Min.   :-2.4   Min.   :-4.3   Min.   :-1.7   Min.   :-2.81
##  1st Qu.: 1.8   1st Qu.:-0.7   1st Qu.:-0.7   1st Qu.:-0.8   1st Qu.:-0.74
##  Median : 2.5   Median :-0.3   Median : 0.0   Median :-0.1   Median : 0.11
##  Mean   : 2.7   Mean   : 0.0   Mean   : 0.0   Mean   : 0.0   Mean   : 0.00
##  3rd Qu.: 3.3   3rd Qu.: 0.7   3rd Qu.: 0.8   3rd Qu.: 0.7   3rd Qu.: 0.81
##  Max.   :15.6   Max.   : 3.3   Max.   : 2.9   Max.   : 3.5   Max.   : 1.84
```

```
hist(data$crime)
```



We can see that all the variables are continuous. Criminal damage rates are approximately normally distributed, although we have a couple of outliers slightly skewing the distribution right. All the other variables are standardised (they have been transformed so they have a mean of 0 and standard deviation equal to 1) so we can compare their relative effect size more easily.

To be precise, before being standardised *unemployment* captured the unemployment rate across MSOAs, *white* represents the percentage of residents who are of white ethnicity, *age* captures the median age, and *collective* is the area weighted average of individual scores of collective efficacy, which was estimated using confirmatory factor analysis based on five items derived from the Metropolitan Police Service Public Attitudes Survey: 'people in this area can be trusted', 'people act with courtesy to each other', 'people take pride in their environment', 'if any young people here are causing trouble, people will tell them off', 'people call the

police if someone is acting suspiciously' and 'if I sensed trouble, I could get help from people who live here'.

In this first exercise we will just seek to approximate the conditional effect attributed to *white*, *collective*, *unemployment*, and *age* on criminal damage. To do so we estimate a linear model with *crime* as the outcome variable and all the other variables in our dataset introduced as explanatory variables.

```
#Modelling damage
naive = lm(crime ~ collective + unemployment + white + age, data=data)
summary(naive)
```

```
##
## Call:
## lm(formula = crime ~ collective + unemployment + white + age,
##     data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.718 -0.735 -0.196  0.521 13.334
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.7055     0.0400   67.64  < 2e-16 ***
## collective    -0.2622     0.0536   -4.89  1.2e-06 ***
## unemployment   0.3251     0.0614    5.30  1.4e-07 ***
## white          0.2934     0.0577    5.09  4.3e-07 ***
## age           -0.2612     0.0571   -4.58  5.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 977 degrees of freedom
## Multiple R-squared:  0.17,   Adjusted R-squared:  0.166
## F-statistic: 49.9 on 4 and 977 DF,  p-value: <2e-16
```

We find that all explanatory variables are statistically significant, and most of them point in the expected direction. Areas with older residents, and higher levels of collective efficacy see less crime, while unemployment has the expected criminogenic effect. The only unexpected effect is that of *white*. We also see that the relative effect size is rather similar across all explanatory variables, which seem to be only weakly associated with crime. For example, areas where the median age is one standard deviation over the London average, ony see 0.26 less criminal damage incidents per 1000 residents and year.

However, we know that police data is heavily flawed. To calibrate our crime variable we can use estimates derived from the Crime Survey for England and Wales (CSEW). Specifically, since the CSEW asks participants whether they have experienced different types of crimes, and whether they reported them to the police, we can estimate the national recording rates for different crimes. For the case of criminal damage we find that recording rates varied from 25% in 2011/12 to 45% in 2018/19 (Pina-Sánchez et al., 2023b). That is, depending on the year, we might see up to three quarters of criminal damage incidents are missing from police statistics.

We can use the recording rate seen in 2011 (the year where most our data comes from) to adjust our *crime* variable and re-estimate our model. To do so we need to keep in mind the functional form of the expected measurement error mechanism at play. In our case, we saw that it makes sense to think of the measurement error present in police data as multiplicative, $Y^* = Y \cdot U$. Hence, if we take $U = 0.25$ then we can rearrange the measurement model and estimate the true level of crime as follows: $Y = Y^*/U$.

```
#Adjusting criminal damage according to the estimated under-recording rate.
data$adj1_crime = data$crime / 0.25
#Re-estimating our model using adjusted criminal damage rates.
```

```r
adj1 = lm(adj1_crime ~ collective + unemployment + white + age, data=data)
adj1
```

```
##
## Call:
## lm(formula = adj1_crime ~ collective + unemployment + white +
##     age, data = data)
##
## Coefficients:
##  (Intercept)    collective  unemployment         white           age
##        10.82         -1.05          1.30          1.17         -1.04
```

```r
#To facilitate comparisons we can put estimates from the naive and adjusted
#models together in a a table.
results = cbind(naive$coefficients, adj1$coefficients)
colnames(results) = c("naive","adj1")
results
```

```
##              naive adj1
## (Intercept)   2.71 10.8
## collective   -0.26 -1.0
## unemployment  0.33  1.3
## white         0.29  1.2
## age          -0.26 -1.0
```

We can now see that the effect of adjusting for the under-recording of criminal damage is to increase the effect size of each of the regression coefficients (including the intercept) proportionally. This helps us see that the actual crime reduction effect associated to *age* or *collective* is much stronger than previously noted.

Now, we do not know for sure that 25% is the true under-recording rate. In fact, the police often detects crimes that are not reported to them, so it is likely that the actual rate is higher than that. So, rather than providing a single adjusted estimate, it would be better if we could report the findings that we obtain after considering a range of recording rates. For example, we could repeat our adjustment considering recording rates ranging from 0.25 to 0.45.

To do so we can repeat the same code that we used above and report findings one by one, or we can automatise that process using a loop. Below we can see how to build such a loop to retrieve the effect of *unemployment* on *crime* for different recording rates.

```r
#Specifying the range of recording rates to be considered.
recording = c(0.25, 0.30, 0.35, 0.40, 0.45)
#Create an empty vector to store the regression coefficients for unemployment.
beta_unemployment = numeric(length(recording))
#Loop over the range of recording rates.
for (i in seq_along(recording)) {
  #Adjusting crime rates.
  data$adj_crime = data$crime / recording[i]
  #Re-estimating the outcome model using adjusted criminal damage rates.
  adj_model = lm(adj_crime ~ collective + unemployment + white + age, data=data)
  #Saving the effect of unemployment.
  beta_unemployment[i] = coef(adj_model)[3]
}
#Printing the effect of unemployment for different recording rates.
adj_results = data.frame(recording = recording, unemployment = beta_unemployment)
print(adj_results)
```

```
##   recording unemployment
```

```
## 1      0.25         1.30
## 2      0.30         1.08
## 3      0.35         0.93
## 4      0.40         0.81
## 5      0.45         0.72
```

We can take these results to convey that areas where *unemployment* is one standard deviation above the London average tend to see 0.7 to 1.3 higher criminal damage incidents per 10000 residents and year, depending on the specific recording rate, and after controlling for median age, ethnic homogeneity, and neighborhood collective efficacy.

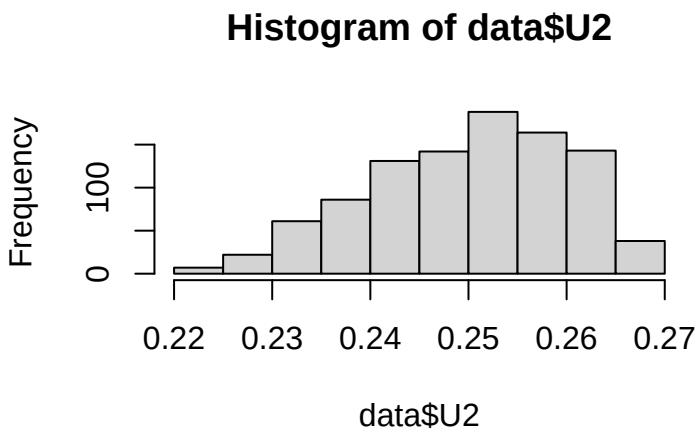
Exercise 2a. Adjusting for differential errors in ethnic homogeneity

Let's now explore in more detail the positive effect of ethnic homogeneity on criminal damage, which we found counter-intuitive since we would expect areas where more white residents live to be wealthier and therefore less prone to crime. According to our naive model, areas where the proportion of white residents is one standard deviation over the London average see an additional 0.29 criminal damage incidents per 1000 residents and year. The effect was even stronger (1.17) after adjusting for an overall recording rate of just 25%.

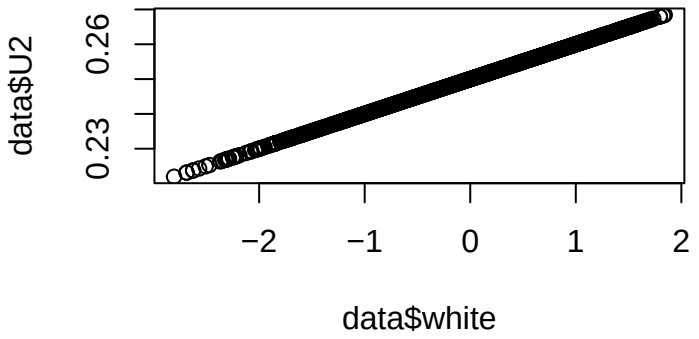
This adjustment improves the robustness of our analysis. However, notice that we are assuming a uniform recording rate across London. That is a questionable assumption, as it is well documented that trust in the police is lower amongst ethnic minority groups (Tyler, 2005). The specific literature exploring crime reporting across ethnic groups is not settled though. Buil-Gil et al. (2021), after estimating the dark figure of crime using the CSEW, shows that the average reporting for different ethnic groups changes significantly year by year. The starkest difference was measured in 2014/15 with seven percentage points between white and black participants, in favour of the former.

To be certain that the positive association between *white* and *crime* is robust we should contemplate different scenarios where we allow for recording rates to differ across ethnic groups. The key here is to generate scenarios that are more plausible than accepting a uniform recording rate across ethnic groups. The simplest way to undertake such assessments is by considering different levels of association between *white* and the error term (U) representing the recording rate. I suggest starting with a weak association in the expected direction (recording rates are higher in neighborhoods with more white residents), which can be formulated using a linear model. We can take the base rate of a 25% recording rate to represent the intercept, and 0.01 as the slope for *white*.

```
#Simulating differential errors.
data$U2 = 0.25 + 0.01*data$white
#Adjusting crime rates according to our selection of differential errors.
data$adj2_crime = data$crime / data$U2
#Notice how the systematic error is not uniform anymore.
hist(data$U2)
```




```
#Recording rates are proportional to the proportion of white residents.
plot(data$white, data$U2)
```



To gain an intuition about whether the recording rates are distinct enough across the distribution of *white* we can examine two key milestones, such as the first and fourth quintiles.

```
#The criminal damage recording rate in the top quintile of white.
data$U2[data$white == quantile(data$white, 0.80)]
```

```
## [1] 0.26 0.26 0.26 0.26 0.26
```

```
#The criminal damage recording rate in the lowest quintile of white.
data$U2[data$white == quantile(data$white, 0.20)]
```

```
## [1] 0.24 0.24 0.24
```

That is, the differential error that we have simulated for the recording rates of criminal damage differs in roughly two percentage points from the lowest to the highest quintile of most ethnically homogeneous areas. If we expected a stronger association we could consider replicating the simulation with a larger value for the effect of *white* on U. For example, we could consider slopes of 0.015 and 0.02.

```r
#Simulating differential errors for a white slope of 0.015
data$U3 = 0.25 + 0.015*data$white
#Adjusting crime rates according to our selection of differential errors.
data$adj3_crime = data$crime / data$U3
#The criminal damage recording rate in the top quintile of white.
data$U3[data$white == quantile(data$white, 0.80)]
```

```
## [1] 0.26 0.26 0.26 0.26 0.26
```

```r
#The recording rate in the lowest quintile.
data$U3[data$white == quantile(data$white, 0.20)]
```

```
## [1] 0.24 0.24 0.24
```

For a slope of *white* equal to 0.015, we have a difference in recording rates across the lowest and top quintile of roughly three percentage points.

```r
#Simulating differential errors for a white slope of 0.02
data$U4 = 0.25 + 0.02*data$white
#Adjusting crime rates according to our selection of differential errors.
data$adj4_crime = data$crime / data$U4
#The criminal damage recording rate in the top quintile of white.
data$U4[data$white == quantile(data$white, 0.80)]
```

```
## [1] 0.27 0.27 0.27 0.27 0.27
```

```r
#The recording rate in the lowest quintile.
data$U4[data$white == quantile(data$white, 0.20)]
```

```
## [1] 0.23 0.23 0.23
```

For a slope of 0.02 we have a difference of roughly four percentage points. All these scenarios seem plausible, so let's see what impact they can have on our findings. Specifically, let's check if the positive association between *white* and *crime* remains after we take into consideration the additional sets of differential errors we have simulated.

```r
#Estimating the adjusted models for each of the sets of differential errors.
adj2 = lm(adj2_crime ~ collective + unemployment + white + age, data=data)
adj3 = lm(adj3_crime ~ collective + unemployment + white + age, data=data)
adj4 = lm(adj4_crime ~ collective + unemployment + white + age, data=data)
#Combining the results for all the adjustments we have considered.
results = cbind(rbind(summary(naive)$coefficients[4,1], summary(naive)$coefficients[4,4]),
                rbind(summary(adj1)$coefficients[4,1], summary(adj1)$coefficients[4,4]),
                rbind(summary(adj2)$coefficients[4,1], summary(adj2)$coefficients[4,4]),
                rbind(summary(adj3)$coefficients[4,1], summary(adj3)$coefficients[4,4]),
                rbind(summary(adj4)$coefficients[4,1], summary(adj4)$coefficients[4,4]))
results_df = as.data.frame(results)  #Convert the results to a data frame.
# Apply formatting to control decimal places.
results_df = format(round(results_df, digits = 3), nsmall = 3)
# Assign column and row names
colnames(results_df) = c("naive", "adj1", "adj2", "adj3", "adj4")
rownames(results_df) = c("coefficient", "p-value")
results_df
```

```
##             naive  adj1  adj2  adj3  adj4
## coefficient 0.293 1.174 0.757 0.536 0.304
## p-value     0.000 0.000 0.001 0.020 0.188
```

This means that, if we take into account the plausible scenario whereby crime recording is lower in areas with

a higher concentration of ethnic minority residents, then we can see how the positive association between *white* and *crime* is lower than we had previously anticipated. In fact, if we see differences in reporting of roughly four percentage points across the highest and lowest quintiles of ethnically homogeneous neighbourhoods, then the association between *white* and *crime* becomes non-significant. Hence, we cannot robustly claim that ethnic homogeneity leads to more crime since this could be the spurious result of white residents showing a higher propensity to reporting crime to the police than ethnic minority residents.

Exercise 2b. Adjusting for differential errors in unemployment

In this exercise you are asked to carry out a similar sensitivity analysis to explore the robustness of the criminogenic effect of *unemployment*. Buil-Gil et al. (2021) indicates that crime reporting rates are slightly lower amongst the unemployed. On average, across the window of observation considered by the authors, this difference does not amount to more than a one percentage point of difference, but it could get up to five percentage points depending on the year.

To undertake this sensitivity analysis decide first on a value for the average recording rate for criminal damage. Then, simulate a range of sets of differential errors that you think could be plausible. Use these simulated errors to adjust the police recorded crime rates in our dataset. Lastly, re-estimate our naive model using different sets of adjusted crime rates. Is the previously noted positive association between *unemployment* and *crime* robust to the expected differential errors? And do you find grounds to suspect that this relationship might have been over or underestimated in the past?

################## EXERCISE ##################

Exercise 2c. Adjusting for differential errors in collective efficacy using RCME

The approach we have followed to simulate differential errors in recording rates provide a good enough approximation. However, we can encounter important limitations when we consider crime types with very low or very high reporting rates. For example, cases of rape tend to be reported to the police in less than 10% of instances, whereas car theft tends be reported in more than 90% of cases (Pina-Sánchez et al., 2023a). In those instances, assuming a linear model to simulate differential errors could easily create recording rates that are lower than zero or higher than one, which represent non-sensical values. To reflect such floor and ceiling effects we recommend using a logit model for the simulation of differential recording rates. We (Pina-Sánchez et al., 2023b) have created an R package, `rcme`, which incorporates such transformation so we can obtain differential recording rates using either odds ratios or risk ratios, which should have the added value of facilitating the interpretation of how plausible those differential errors are.

In this exercise we are going to practice using `rcme` to explore the robustness of the crime reduction effect attributed to *collective*. Collective efficacy is a widely studied concept in Criminology, understood as the combination of ties between local residents with a willingness to intervene to prevent crimes (Mazerolle et al., 2010; Sampson et al., 1997). The literature is clear in that low levels of collective efficacy are seen as a key precursor of crime. However, there are theoretical grounds to expect that crime reporting is positively associated with collective efficacy, since social trust and willingness to collaborate with the police are specific items used in the measurement of collective efficacy. This intuition has been empirically corroborated in the literature. For example, at the individual level, Hart and Colavito (2011) reported a positive association between collective efficacy and crime reporting amongst US college students, while in England and Wales, Brunton-Smith et al. (2022) report a similar positive association between collective efficacy and crime recording at the Community Safety Partnership level.

To be able to use `rcme` you need to install it first, to do so you also need to install the `devtools` package if you do not have it already. You can install both of those packages with the two lines of code commented out below. If you are asedk to press enter to continue the installation, then do so. Once the packages are installed we can call the `rcme` library.

```
#install.packages("devtools")
#devtools::install_github("RecountingCrime/rcme")
library(rcme)
```

To undertake this sensitivity analysis, where *crime* is considered the outcome variable, we will use the `rcme_out` command (when *crime* is introduced as an explanatory variable you should use `rcme_ind`). To use `rcme_out` we need to provide the following information: i) the type of outcome model we are exploring, `formula`; ii) the `data` we are using; iii) the `focal_variable`, i.e. the variable which effect on crime we are investigating; iv) the expected average recording rate, `R`, or a range of plausible recording rates; and optionally, v) the type of differential errors, `D`, expressed as an odds ratio or as a range of them; and vi) whether *crime* has been log-transformed, `log_var`.

For our case we will consider recording rates ranging from 25% to 45% and differential errors ranging from an odds ratio of 0.95 (i.e. areas with a one standard deviation of *collective* one standard deviation higher than the London average show 5% lower odds of recording crime) to an odds ratio of 1.10 (i.e. areas with a one standard deviation of *collective* one standard deviation higher than the London average show 10% higher odds of recording crime). We can pass that information to the `rcme_out` command as shown below. This will then: i) simulate the different errors for our combination of average recording rates, and differential errors; ii) adjust *crime* accordingly (assuming a multiplicative measurement error model); iii) re-estimate the outcome model for each of the adjusted crime rates created; and iv) report the effect of *collective* under each of the scenarios considered.

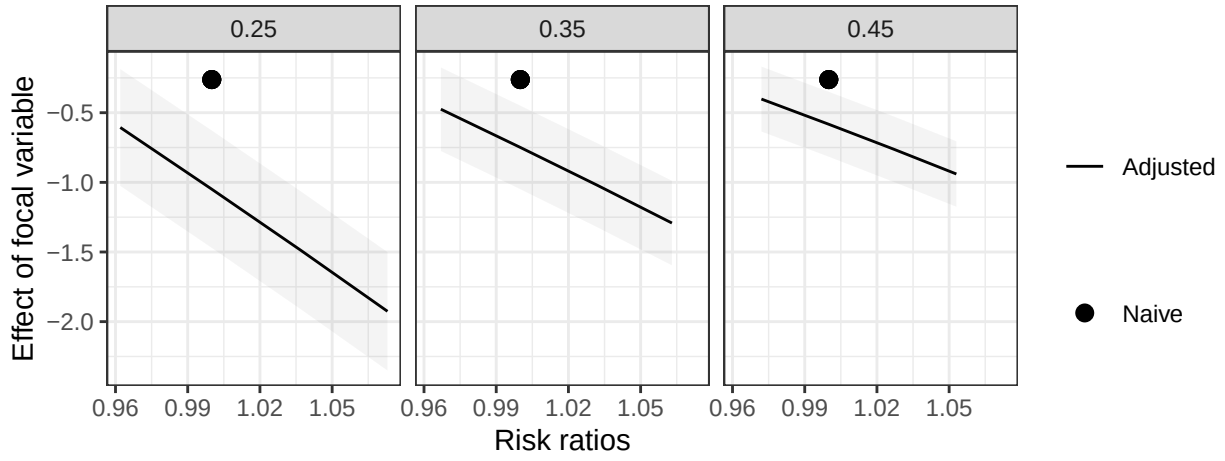
```
robust_collective = rcme_out(
  formula = "crime ~ collective + unemployment + white + age",
  data = data,
  focal_variable = "collective",
  R = c(0.25, 0.35, 0.45),
  D = c(0.95, 1, 1.05, 1.10))
robust_collective
```

```
## $sim_result
##       R    D log_var   rr focal_variable   SE
## 1  0.25 0.95   FALSE 0.96          -0.61 0.21
## 2  0.35 0.95   FALSE 0.97          -0.47 0.15
## 3  0.45 0.95   FALSE 0.97          -0.40 0.12
## 4  0.25 1.00   FALSE 1.00          -1.05 0.21
## 5  0.35 1.00   FALSE 1.00          -0.75 0.15
## 6  0.45 1.00   FALSE 1.00          -0.58 0.12
## 7  0.25 1.05   FALSE 1.04          -1.49 0.22
## 8  0.35 1.05   FALSE 1.03          -1.02 0.15
## 9  0.45 1.05   FALSE 1.03          -0.76 0.12
## 10 0.25 1.10   FALSE 1.07          -1.93 0.22
## 11 0.35 1.10   FALSE 1.06          -1.29 0.15
## 12 0.45 1.10   FALSE 1.05          -0.94 0.12
##
## $naive
##
## Call:
## lm(formula = paste0(outcome, " ~ ", paste0(predictors, collapse = " + ")),
##     data = data)
##
## Coefficients:
##  (Intercept)    collective  unemployment         white           age
##        2.706        -0.262         0.325         0.293        -0.261
##
##
## $focal_variable
## [1] "collective"
```

We can retrieve the estimated regression coefficient for *collective* (under `focal_variable`) and its associated standard error (`SE`), for each of the combinations of average recording rates and differential errors. In addition, to facilitate interpreting the plausibility of the differential errors invoked, we are also given the equivalent association in terms of risk ratios (`rr`). Risk ratios represent the ratio of probabilities, which is more intuitive than the ratio of odds. For example, we can interpret the last of the scenarios reported above as the average recording rate for criminal damage assumed to be 45%, but this is 5% higher (47%) in areas where *collective* is one standard deviation above the average.

Ultimately, when we are considering a range of parameters for the average recording rate and potential differential errors, their impact can be interpreted best visually, which we can generate using `rme_sim_plot`. Below we do so while requesting to include the 95% confidence intervals (`ci`), the point estimate for the effect of *collective* on *crime* under the `naive` model (i.e. ignoring measurement error), and differential errors expressed as risk ratios (`rr`).

```
rcme_sim_plot(robust_collective, ci = T, naive = T, rr = T)
```



We can see how the crime reduction effect of *collective* is severely under-estimated, with the magnitude of that bias being proportional to the extent of under-recording and the differential errors. In fact, even if we consider a 0.95 risk ratio as our value of differential errors, that is, even if we consider that recording rates are lower in areas where collective efficacy is higher (an assumption that contradicts the available literature), we can still see that the effect of *collective* remains underestimated.

We can therefore use `rcme` to facilitate the sensitivity analysis of regression models that include police recorded crime data as either their outcome or as an explanatory variable, considering any range of recording rates and differential errors. This is a highly flexible approach, we have illustrated how it works for the case of a linear outcome model, but it can also be used for any other type of regression model (changing the `formula` option in `rcme_out`). In addition, even though `rcme` was built with the goal of facilitate measurement error adjustments in police data, it could also be used for the adjustment of any other variable subject to systematic multiplicative errors, such as retrospective reports of the number or duration of spells of unemployment (Pina-Sánchez et al., 2016), personal income (Glewwe, 2007), or recalled dates of lifecourse milestones, like age at menarche (Pickles et al., 1996).

To learn more about `rcme` see here, https://github.com/RecountingCrime/rcme.

Exercise 3. Failing to adjust for random errors

In this workshop we have practiced simulating multiplicative systematic errors, such as those that can be expected from police recorded crime rates. We have also simulated differential errors to reflect varying crime rates for different values of continuous variables like ethnic homogeneity or collective efficacy. These are only a couple of examples of the types of measurement errors that could be simulated to undertake similar

sensitivity analyses (see Gallop and Weschle, 2019). There is however a common form of measurement error that we cannot strip away simply through simulations, that is random measurement error.

Unlike for the case of systematic errors, when we have a variable prone to random errors, we cannot adjust it directly because we do not know (not even approximately) how each value is affected. So, even if we know, say, the reliability ratio of the observed variable, which can be used to simulate a similar type of measurement error process to the one present in that variable, we cannot use that simulated noise to recover the true variable. Below we demonstrate this limitation empirically.

Let's simulate two continuous variables X and Y, each following a standard normal distribution, and let's assume that the two variables are linearly associated with X having an effect of 0.5 on Y.

```
#The true data.
X = rnorm(10000, 0, 1)
#Y is caused by X.
Y = 0.5*X + rnorm(10000, 0, 1)
```

Let's now take X to be unobserved, instead we observe X^*, which is affected by classical measurement error, U.

```
#Simulating classical errors.
U1 = rnorm(10000, 0, 0.5)
#The observed version of X, affected by classical errors.
Xstar = X+U1
#Notice how the classical errors inflate the variability of X, as expected.
sd(X)
```

```
## [1] 1
```

```
sd(Xstar)
```

```
## [1] 1.1
```

```
#We can calculate the reliability ratio or Xstar as follows.
var(X)/var(Xstar)  #A reliability ratio of 0.8.
```

```
## [1] 0.8
```

Now that we have the true and the observed versions of X we can assess the impact that classical measurement error will have on the causal effect of X on Y. In the lecture we saw how for this simple case of simple linear regression with the explanatory variable affected by random errors we should expect an attenuation effect inversely proportional to the reliability ratio of the error prone variable.

```
#The true model.
true = lm(Y~X)
true
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##    -0.00487      0.50600
```

```
#The naive model.
naive = lm(Y~Xstar)
naive
```

```
##
## Call:
```

```
## lm(formula = Y ~ Xstar)
##
## Coefficients:
## (Intercept)        Xstar
##    -0.00234      0.40629
```

```
#We observe roughly the 20% attenuation that we expected.
```

Crucially, we cannot retrieve the true effect of X on Y by adjusting X^*, even if we know the characteristics of the measurement error mechanism, namely, random errors with a standard deviation of 0.5. Let's check that is the case.

```
#Simulating a random measurement error process like the one that affects Xstar.
U2 = rnorm(10000, 0, 0.5)
#Adjusting the observed variable accoridng to an additive measurement error model.
Xadj = Xstar - U2
sd(X) #The sd of the true variable.
```

```
## [1] 1
```

```
sd(Xstar) #The sd of the observed variable prone to classical errors.
```

```
## [1] 1.1
```

```
sd(Xadj) #The sd of the adjusted variable.
```

```
## [1] 1.2
```

```
#Oh dear, we have made it worse.
var(X)/var(Xadj) #The reliability ratio of the observed variable
```

```
## [1] 0.67
```

```
var(X)/var(Xadj) #Reliability ratio of the adjusted variable.
```

```
## [1] 0.67
```

```
#Rather than adjusting the measurement error we have created more of it.
adj = lm(Y~Xadj)
#Let's check if we have adjusted the effect of X on Y.
true$coefficients[2]
```

```
##    X
## 0.51
```

```
naive$coefficients[2]
```

```
## Xstar
##  0.41
```

```
adj$coefficients[2]
```

```
## Xadj
## 0.34
```

```
#The attenuation effect is even lower than previously.
#In essence we are adding more noise, and making things worse.
```

This incapacity to adjust for random errors represents the main limitation of the type of simulations-based adjustments that we have practised here. In the next practical we will see how we can still use simulations to indirectly adjust for random errors using SIMEX. This involves changing the strategy so we do not not seek to adjust the observed variable directly, but rather we estimate the impact that measurement error have on

our causal estimate of interest through simulating increasing levels of measurement error, and extrapolate to find the true estimate, when measurement error is not present.

References

Brunton-Smith, I., Buil-Gil, D., Pina-Sánchez, J., Cernat, A. and Moretti, A. (2022). 'Using synthetic crime data to understand patterns of police under-counting at the local level'. In: *The Crime Data Handbook*. Policy Press.

Buil-Gil, D., Medina, J., and Shlomo, N. (2021). 'Measuring the dark figure of crime in geographic areas: Small area estimation from the crime survey for England and Wales'. *The British Journal of Criminology*, 61(2), 364-388.

Gallop, M., and Weschle, S. (2019). 'Assessing the impact of non-random measurement error on inference: a sensitivity analysis approach'. *Political Science Research and Methods*, 7(2), 367-384.

Glewwe, P. (2007). 'Measurement error bias in estimates of income and income growth among the poor: Analytical results and a correction formula'. *Economic Development and Cultural Change*, 56, 163–189.

Hart, T. C., and Colavito, V. (2011). 'College student victims and reporting crime to the police: The influence of collective efficacy'. *Criminology, Criminal Justice, Law & Society*, 12(3), 1–19.

Mazerolle, L., Wickes, R., and McBroom, J. (2010). 'Community variations in violence: The role of social ties and collective efficacy in comparative context'. *Journal of Research in Crime and Delinquency*, 47, 3–30.

Pickles, A., Pickering, K., and Taylor, C. (1996). 'Reconciling recalled dates of developmental milestones, events and transitions: A mixed generalized linear model with random mean and variance functions'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(2), 225–234.

Pina-Sánchez, J., Brunton-Smith, I., Buil-Gil, D., and Cernat, A. (2023b). 'Exploring the impact of measurement error in police recorded crime rates through sensitivity analysis'. *Crime Science*, 12(1), 14.

Pina-Sánchez, J., Buil-Gil, D., Brunton-Smith, I., and Cernat, A. (2023a). 'The impact of measurement error in regression models using police recorded crime rates'. *Journal of Quantitative Criminology*, 39(4), 975-1002.

Pina-Sánchez, J., Koskinen, J., and Plewis, I. (2014). 'Measurement error in retrospective work histories'. *Survey Research Methods*, 8(1), 43-55.

Sampson, R. J., Raudenbush, S. W., and Earls, F. (1997). 'Neighborhoods and violent crime: A multilevel study of collective efficacy'. *Science*, 277(5328), 918–924.

Tarling, R., and Morris, K. (2010). 'Reporting crime to the police'. *The British Journal of Criminology*, 50(3), 474-490.

Tyler, T. (2005). 'Policing in black and white: Ethnic group differences in trust and confidence in the police'. *Police Quarterly*, 8(3), 322-342.