Assignment 2

Your name

2023-11-24

You will be asked to perform some basic data manipulation and analysis tasks. You will likely want to use some functions from the tidyverse package that we’ve focused on in the course. The data you’ll be working with are available at this URL:

https://clanfear.github.io/ioc\_iqa/\_data/neighb\_crime.csv

You’ll want to load them into R (neighb\_crime would be a good name) as we’ve done before in class. You’ll probably want to take a quick look at them to get an idea for what is there. These data describe 343 neighborhoods in a major city observed over a 10 year period. It contains the following measures:

* robbery: A count of the number of robberies occurring in the neighborhood within the 10 year study period. Units are individual robbery incidents.
* disadvantage: An index of concentrated disadvantage (e.g., unemployment, poverty). Units are standard deviations and higher values indicate higher disadvantage. e.g., a neighborhood with disadvantage equal to 1 is 1 standard deviation higher in disadvantage than a neighborhood with disadvantage equal to 0.
* instability: An index of residential instability (e.g., home ownership, recent moves). Units are standard deviations and higher values indicate higher instability.
* heterogeneity: An index of ethnic heterogeneity (e.g., percent foreign-born, ethnic diversity). Units are standard deviations and higher values indicate higher heterogeneity
* control: An index of informal social control capacity. Units are standard deviations and higher values indicate higher social control.
* opportunity: A three-level ordinal variable capturing the density of features of the built environment that provide criminal opportunities (e.g., abandoned buildings, street lighting, commercial destinations).

When interpreting the indices (i.e., control, disadvantage, instability, and heterogeneity), you can just say, for example, “a one unit higher value of control is associated with more robberies in a neighborhood”.

You may write answers as normal text in this Word doc below each question. Alternatively, if you’d prefer to make a second document with only your answers (or an RMarkdown / Quarto doc for advanced students), go ahead. You will upload this to the assignment on Moodle. For questions that ask for some output from R, like a table or plot, you can insert them into this document however works best for you. Copy-pasting output from the console is fine, so long as it is passably formatted: format pasted code (but not written text) with a fixed-width/monospace font like Courier or Monaco. Cropped screenshots are fine for console output or plots (only show the relevant parts). Written answers should be brief and to the point!

Also include any code used to solve these problems in a separate R script file (a .R file, like what we use in class time) or in chunks in your Quarto/RMarkdown doc. Divide this code into sections using comments. You can get started with [this template here](https://clanfear.github.io/ioc_iqa/_assignments/assignment-2.R) if you would like some guidance. Avoid writing a lot of code if possible—most problems can be solved with just a few lines of code using what we’ve learned in lecture—but any amount of code that solves the problem you have is good enough!

# Fundamentals

## Correlation

Calculate the correlation between robbery and control, then calculate the correlation between robbery and disadvantage. In a single sentence of plain language, explain what these values mean.

## Linear Model

### Estimates

Estimate a linear model where robbery is predicted by control, opportunity, disadvantage, instability, and heterogeneity. Show the summary output here too—you’ll want it for the next question.

### Equation

Write out the full regression equation from this model (e.g., ). For brevity, you can round the coefficients to 2 decimal places and abbreviate words, as long as it is still clear.

### Prediction

Based on this model, how many robberies would we expect to see in a neighbourhood with “high” opportunity, control at -1, disadvantage at 1, and both heterogeneity and instability at 0.

### R-squared

In 1 or 2 sentences, what is the value for this model and how would you interpret it?

### Interpretation

In 1 or 2 sentences each, how would you interpret (1) the coefficient for instability and (2) the coefficient for “low” opportunity? Include their statistical significance in your interpretation.

## Histograms

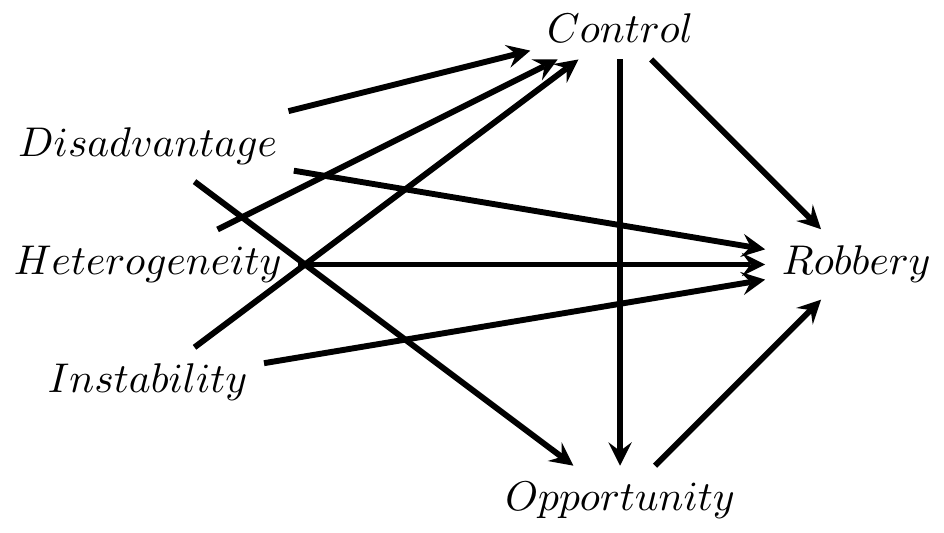
Create a histogram or density plot of the robbery variable. What do you notice about this variable?

Next, repeat your histogram or density plot but use the **log** of the robbery variable. How does this compare to the prior plot?

# Moderate Problems

## DAG to model

Consider the DAG below that depicts a model of social disorganization theory with criminal opportunity added.



Compare your linear model from question 1.2 to the DAG. Based on the statistical significance of the coefficients, do our model results agree with the theory in the DAG? That is, do we find evidence that all the paths on the DAG that lead into robbery exist? Why or why not?

## Log-level Model

We might consider using the **log** of robbery as our outcome instead. Rewrite the linear model from 1.2 to predict the log of robbery using the other covariates. Is it possible to use anova() to compare this new log-level model to the previous linear one?

## Quadratic term

A substantial literature in community criminology suggests disadvantage has a non-linear effect on crime: Once disadvantage gets relatively high, further disadvantage doesn’t contribute much to crime. Add a **quadratic polynomial** functional form of disadvantage to your model from 2.2 (or 1.2 if you didn’t know how to do 2.2).

## Model comparison

Evaluate whether there is statistical evidence that you should use the quadratic functional form of disadvantage instead of the linear term alone (i.e., does it fit better than it did before?).

# Harder Problems

## Interaction plot

Create a scatterplot with control on the x-axis and the **log** of robbery on the y-axis. Include different coloured regression lines for each level of opportunity. In a sentence or two, what do you see in this plot?

## Interaction Model

Based on the plots from 3.1, we might expect there is an **interaction** between control and opportunity. That is, we think the effect of control on robbery might be different at different levels of opportunity. Take your linear model from 2.3—with the logged outcome and quadratic disadvantage—and add an interaction between control and opportunity.

Does the model with the interaction term fit better than the one without? Use a statistical test to compare them and interpret the results of the test.

## Interpretation

This is a complex model—it has a logged outcome and an interaction. Rather than interpreting the coefficients numerically, explain them intuitively. Specifically, at what level of opportunity does control have the strongest association with the log of robbery, and at what level of opportunity does control have the weakest association with the log of robbery?

# How’s it goin?

In anything from one word to a paragraph, let me know how you’re feeling now that the course has come to an end. Were you overwhelmed? Did it get better or worse over the term? No problem at all? Are there topics that leave you confused? Let me know if there are things that might make the course work better for others in the future. Also let me know if there are topics you would like to cover in the future, in case I can set up some workshops.