What causes what?

September 14, 2023

[99]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import math
import matplotlib.pyplot as plt
import statsmodels.api as sm
from sklearn.model_selection import train_test_split, GridSearchCV, KFold,␣
 ↪cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.metrics import classification_report
```

[100]:
```
# Creating a function to print output in green and bold
# ANSI escape code for green color and bold font
GREEN_BOLD = '\033[1;32m'

# ANSI escape code to reset colors and font style
RESET = '\033[0m'

def print_green_bold(*args):
    text = ' '.join(str(arg) for arg in args)
    print(GREEN_BOLD + text + RESET)
```

1 What causes what??

Listen to this podcast: http://www.npr.org/blogs/money/2013/04/23/178635250/episode-453-what-causes-what 1. Why can't I just get data from a few different cities and run the regression of "Crime"on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city)

Crime levels in a city are influenced by a multitude of factors and cannot simply be explained away by the number of cops in the city. We can look at crime levels and number of cops in a city in an attempt to understand if there is a correlation between the two but simply running a regression will not be meaningful as correlation does not imply causation.

2. How were the researchers from UPENN able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2" below.

Effect of Police on Crime

TABLE 2

Total Daily Crime Decreases on High-Alert Days

	(1)	(2)
High Alert	-7.316*	-6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

They started looking for examples of areas where the number of cops were very high but for reasons unrelated to crime.This process brought them to the terrorism alert system. Since Washington DC is likely to be heavily targeted by terrorists, the government has an alert system to gauge threats. When the threat level hits a certain point (say, orange on the scale), more police officers are deployed to various areas across the city. This is completely unrelated to street crime. Table 2 contains the results of two regression models run by the researchers. They regressed the total daily crime in DC on just the high alert dummy in the first model and both the high alert dummy and log of metro ridership in the second model. In both cases they found that the coefficient for the high alert dummy was negative and statistically significant. This implies that the larger number of cops in the street on high alert days does indeed negatively impact the amount of street crime.

3. Why did they have to control for METRO ridership? What was that trying to capture?

The researchers hypothesized that the on high alert days, there could be lesser tourists on the streets i.e. a reduced number of potential victims and subsequently a reduced amount of crime. In order to test this theory, they looked at whether metro ridership (a potential measure of the number of tourists out and about) reduced on high alert days but ultimately found that metro ridership remained fairly stable and showed no downward spikes during high alert days.

In Table 2, they also used metro ridership as one of the regressors for daily crime levels and found that it was statistically significant with a positive coefficient. This means that a larger number of tourists does indeed correspond to a higher level of crime. But since they could not prove that there were lesser tourists on the streets on high alert days, they cannot attribute the lower levels of crime on those days to this variable.

4. Just focus on the first column of Table 4. Can you describe the model being estimated here? What is the conclusion?

TABLE 4

Reduction in Crime on High-Alert Days: Concentration on the National Mall

Coefficient Type	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert # District 1	-2.621** (.044)	-2.621* (1.19)	-2.621* (1.225)

Coefficient Type	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert # Other Districts	-.571 (.455)	-.571 (.366)	-.571 (.364)
Log(midday ridership)	2.477* (.364)	2.477** (.522)	2.477** (.527)
Constant	-11.058** (4.211)	-11.058 (5.87)	-11.058 ” (5.923)

In this model, the daily crime rate in DC is being regressed on three variables : - High Alert * District 1 : This is an interaction term that tries to gauge whether having high alert in district 1 has a significant impact on daily crime - High Alert * Other Districts : This variable tries to capture whether having a high alert in districts other than district one has a significant impact on daily crime - Log(midday ridership) : This variable is used as a potential measure of the number of tourists/potential victims on the streets

If we focus on the first column as recommended, we observe the following: - The High Alert * District 1 variable has a negative cofficient which is statistically significant at the 1% level - The High Alert * Other Districts variable has a negative coefficient but is not statistically significant (Note that the coefficient here is much smaller in absolute terms than for High Alert * District 1) - The log(midday ridership) variable has a positive coefficient and is statistically significant at the 5% level

This means that overall daily crime level in DC is significantly impacted by the presence of a high alert in district 1 but the same does not hold true for a high alert in other districts. Thus, having a high alert and therefore more cops in districts other than District 1 does not really reduce crime. The results also demonstrate that a larger number of people out and about does increase the number of potential victims and therefore increases the incidence of crime.

[ ]: