Exploration of a Linear Relationship between Burglary and Larceny Rates

ABSTRACT

Though the media focus on crime may lead one to believe that crime rates are on the rise, the opposite is seemingly the case. Crime rates as a whole have generally dropped with time. However, in the case of burglaries, the monetary impact of burglaries has increased. Burglary tends to be a higher class crime than the misdemeanor equivalent of larceny, so it is conceivable that law enforcement has greater influence over larceny rates than burglary rates. As such, this paper leverages crime data from 1960-2014 to look into the relationship between burglary rates and larceny rates. Specifically, simple linear regression and statistical inference with wild bootstrap are used to describe the significant relationship between burglary rates and larceny rates in the U.S.

Introduction

Crime has always been a unique obsession for the American people. Whether it be nightly news reports or grandiose television programs, crime always holds a dedicated spot in mass media. With such prominence, this programming helps fuel rhetoric that proposes that crime rates have only increased with time. However, a study by the U.S. Department of Justice implies that burglary rates have "decreased 56% from 1994 to 2011" [2]. This statistic alone seems to suggest that burglary is not necessarily the most pressing crime on which to focus. Yet, it is estimated that only 58% of burglaries are reported and the median value of stolen items has risen from $389 to $600 annually [2]. Thus, while the decline in burglary rates seems promising, other metrics serve as cautionary reminders that complacency regarding this decline could have damaging repercussions.

The U.S. Department of Justice study only looks at burglary in isolation, which may not provide enough insight for decision-makers who are generally interested in property crime. Burglary and larceny are two of the primary crime types which compose property crime. Burglary tends to be a more severe, yet rarer, crime, whereas larceny is typically just a misdemeanor but is more prominent. Knowing this distinction, one may imagine that larceny rates are easier to influence from a law enforcement perspective, as larceny comprises much smaller violations of the law. Thus, the focus of my research is to explore a prospective connection between burglary and larceny rates. Primarily, is there a significant linear relationship between burglary rates and larceny rates? Indeed, such a relationship between larceny and burglary could provide justification for policies that look to control burglary by targeting larceny crime instead. While the U.S. Department of Justice's study is a report which reflects on changes in burglary, my study looks to use such data to analyze the connection between these two distinct forms of crime.

Analysis

The data for my study was obtained from the U.S. Department of Justice's UCR data tool [1]. I obtained property crime rate data from 1960 - 2014 for all 50 U.S. states and the District of Colombia. This information was compiled into a single long-format table for analysis and is available for reanalysis online [3]. Within this table, there are 2798 cases, which approximately accounts for the 55 years for all 50 states and the District of Columbia. The only modification made to this data was to capitalize all of the "State" variable's values to keep the format consistent across the entirety of the dataset. For this analysis, I will only consider the burglary rate and larceny rate variables, which track the respective crime rates per 100,000 in the population.

Given the question of interest, I fit a simple linear regression model in burglary rates and larceny rates. In this model, burglary rate is the response and larceny rate is the predictor. This choice is made because we are interested in justifying the proactive policies that look to control burglary by targeting larceny crime. A scatter plot found in the appendix helps visualize this relationship and supports the exploration of a linear relationship between these variables.

Before we consider the fitted model, we will first define some formal variables:

B: reported burglary rate (per 100,000 population) in a given state during a given year

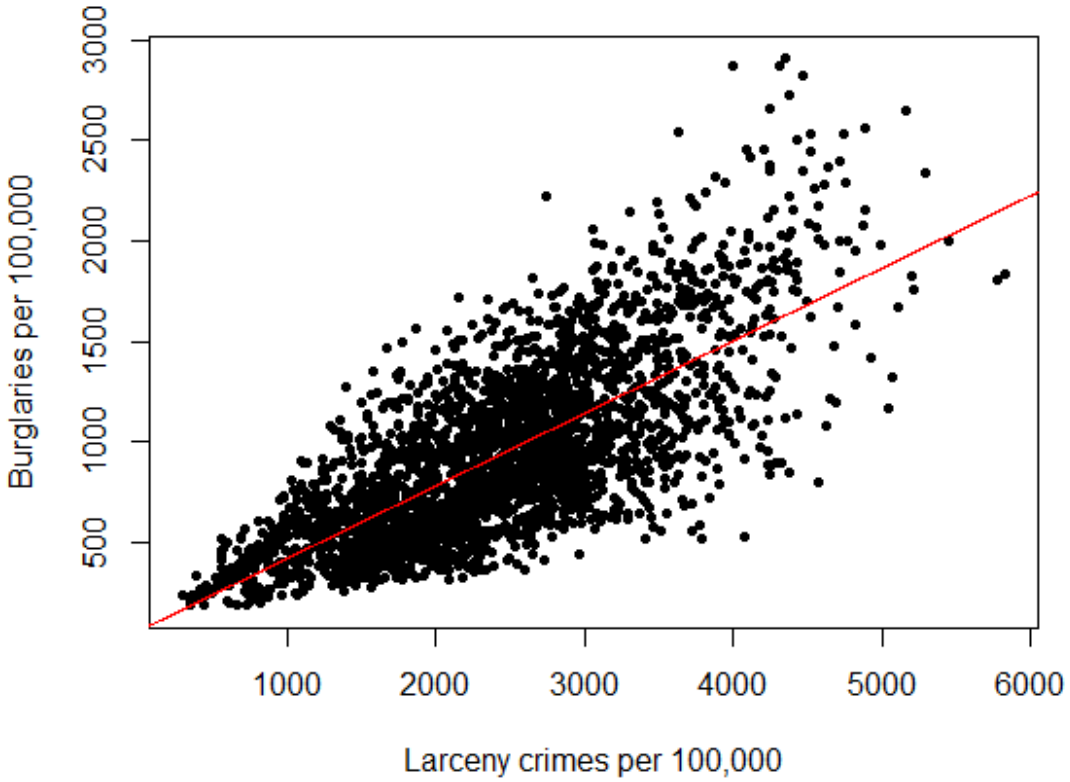
L: reported larceny rate (per 100,000 population) in a given state during a given year

Using least squares estimation, we obtain the following linear model,

$$\hat{\mathrm{B}} = 57.31 + 0.36(\mathrm{L})$$

where $\hat{\mathrm{B}}$ is the conditional mean burglary rate given larceny rate.

Figure 1: Observed U.S. Burglary vs. Larceny rates with the fitted linear model



Our observed property crime rate data is plotted with the fitted linear model above in Figure 1. Notice that the linear model does seem to be an appropriate form to describe the relationship between these two variables. However, we see that there are plenty of points that deviate fairly heavily from this line. One goal in predictive modeling is to create a model which explains a high proportion of variability in the response -- here burglary rate. The R^2 of this particular model is 0.551, which indicates approximately half of the variability in burglary rate is explained by a linear model in larceny rate.

This variability may be traced to an important aspect of this data structure that we have glossed over. That is, the data likely contains much more complicated spatial and temporal relationships than are addressed here. These ideas are discussed further in the conclusions. Note that citing these notions is not to say that our model is useless, but rather that there is room for further work that could improve this model. Now, we need to address assumptions for statistical inference.

Detailed discussion of assumptions for statistical inference is provided in the appendix. Of key importance to note, if we may cautiously assume that the errors are independent, it is not reasonable to use the t-distribution when drawing inference over our slope coefficient. Figure 1

suggests the finding that it is not reasonable to assume errors are identically distributed. The heteroscedasticity is fairly obvious from the fan-shaped pattern in our plotted data. Thus, we choose to use the wild bootstrap [4] to overcome this and create a confidence interval for our slope coefficient.

Our main interpretations focus on the proportional relationship between burglary and larceny rates (especially since the intercept is not present in our observed data). In context, we are 95% confident that whenever there is an increase of 100 larceny crimes per 100,000 people, we expect there to be a corresponding increase of between 34.7 burglaries and 37.5 burglaries per 100,000 people on the average. Thus, we have found a statistically significant linear relationship between burglary and larceny rates. The next reasonable step in this analysis would be to create 95% confidence and prediction regions for this model, which would be much more useful in decision-making. This functionality is currently being added to the bootstrap functions used and thus may be implemented in the near-future.

Conclusions

Our results were not entirely unexpected, as they depicted a moderate positive linear relationship between burglaries and larceny crimes. Though this model does not perfectly relate burglaries and larceny crimes, there is a substantive enough connection that the results are still of some use. By looking at crime on a more granular level than, for example, property crime totals, the results are more actionable for law enforcement agencies. For these agencies, it is imaginable that larceny, as a misdemeanor, is easier to influence than burglary, a more severe crime. Thus, when agencies can see what reductions in larceny crime could be expected to do to burglary rates, then it is easier to justify action. This model alone is not substantive enough to incite such policy changes, but this form of analysis is easily applicable to other crimes and thus has greater use as a pattern by which to explore the connections between other, perhaps more pressing, forms of crime.

The most obvious augmentation to this model is to apply this methodology to other pairings of crime categories and see what correlations can be found. What that type of work amounts to is simply creating new models in a similar vein to this one, which is a bit tangential to improving this model. Hence, it might be more interesting to continue this work by delving into the possible influence that spatial factors had on the results. Exploring this model at a regional or state level would likely produce wildly different results, which could be compared across regions and to our overall model for unique trends. Such an analysis would require the use of multiple linear regression and would likely be the immediate step I would take next. Finally, considering other types of crime, such as those which comprise violent crime, and time as variables in the linear model may yield significant, interesting relationships.

References

[1] U.S. Federal Bureau of Investigation. (2017). "Uniform Crime Reporting Statistics." Retrieved from https://www.ucrdatatool.gov/

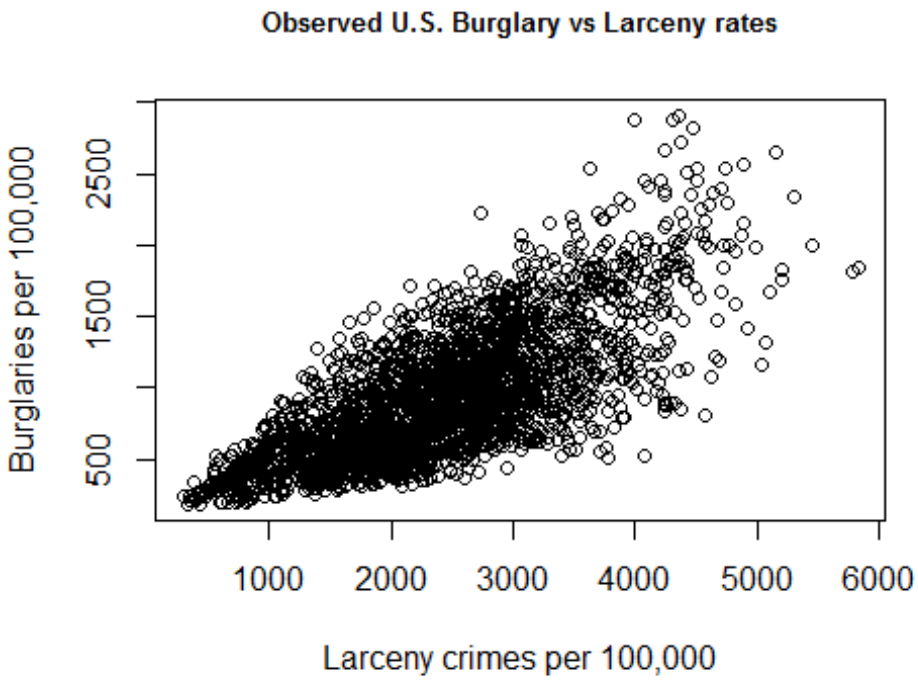
[2] Walters, J., Moore, A., Berzofsky, M. and Langton, L. (2013). *Household Burglary, 1994-2011*. Bureau of Justice Statistics.

[3] Heyman, M., RHIT Data, (2017), GitHub repository, https://github.com/meganheyman/RHIT-Data

[4] Wu, C. F. J. (1986). "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis (with discussion)." Annals of Statistics, 14:1295.

Appendix

Plot of Burglary and Larceny Rates

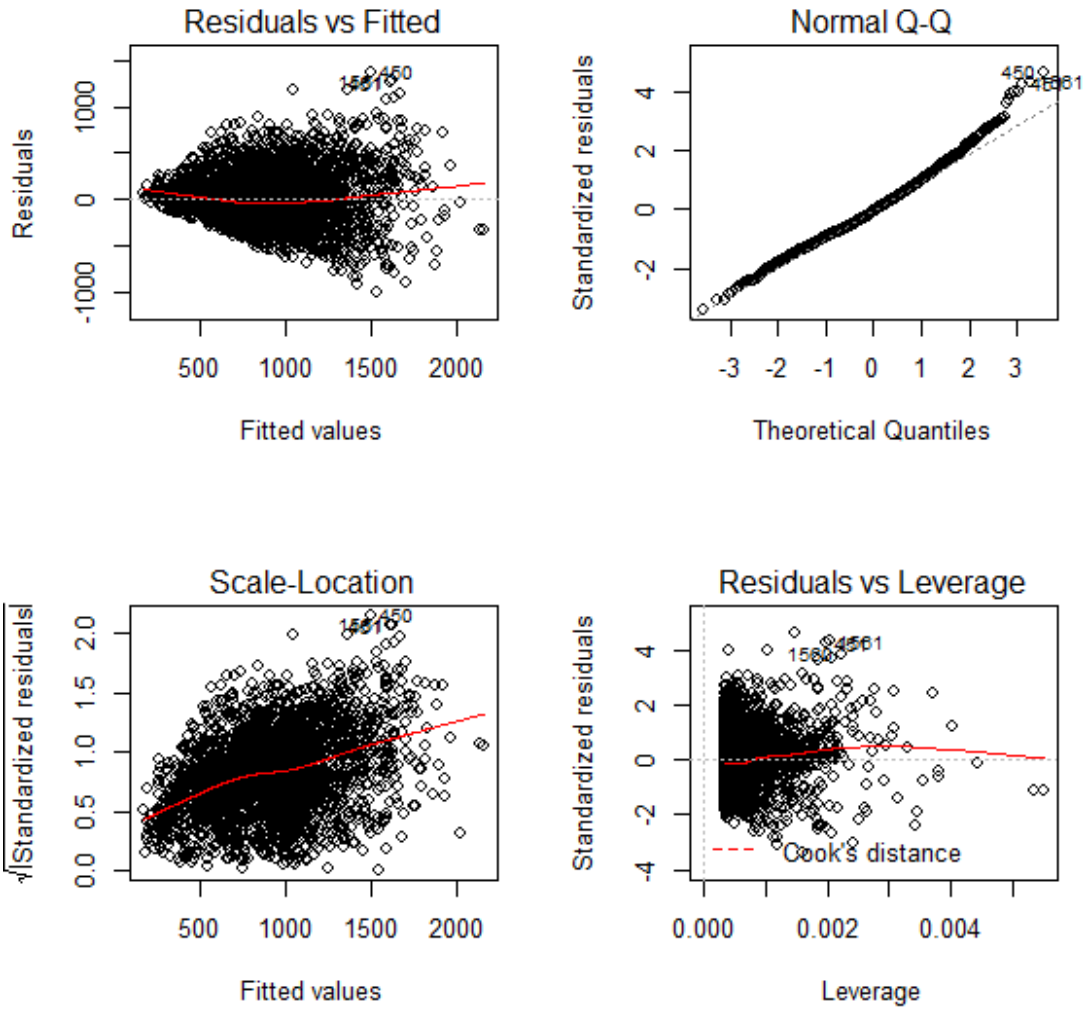


Linear Model Summary

```
##
## Call:
## lm(formula = mergedData$Burglary.rate ~ mergedData$Larceny.theft.rate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1005.74  -205.19   -20.47   178.85  1368.36
```

```
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   57.310460  15.676261   3.656 0.000261 ***
## mergedData$Larceny.theft.rate  0.361175   0.006156  58.666  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 296.7 on 2798 degrees of freedom
## Multiple R-squared:  0.5516, Adjusted R-squared:  0.5514
## F-statistic:  3442 on 1 and 2798 DF,  p-value: < 2.2e-16
```

Discussion of Assumptions



We use the residual diagnostic plots to discuss the extent to which assumptions for statistical inference hold below:

1. Errors are independent.

Research into the data collection methodology indicates that this data was collected by the FBI from agencies who voluntarily participated in crime statistic collection. This voluntary participation alone suggests that we should be cautious with interpretations to the general population, but does not necessarily eschew the possibility of independent errors. However, this methodology also reports for each state on a yearly basis. Now, knowing these facts, it is entirely likely that agencies within a state may opt out on a year to year basis, leading to shifts in variability that could very well be time-dependent. Given the limited details beyond the above on the collection methodology, it is difficult to come to a conclusion regarding error independence. My inclination is to cautiously say that the errors are independent, given the somewhat haphazard methodology of reporting, assuming that agencies have the power to participate or not as they see fit. Further work should attempt to incorporate the state and temporal data.

2. Errors have mean 0.

We see no obvious trends in the residuals vs. fitted plot. So, given the values displayed in said plot, it's fairly reasonable to work under the impression that this assumption holds.

3. Errors are identically distributed.

The residuals vs. fits plot has a clear fan, which indicates that there is evidence of non-constant variance in the errors.

4. Errors come from a normal population.

Since the third assumption is unreasonable, this assumption is not discussed, since the construction of the normal Q-Q plot depends upon identical distribution of errors.

Given that only the first two assumptions are fairly reasonable, wild bootstrap is our method of choice for statistical inference.

Wild Bootstrap Output

```
#########################################################
## Read bootstrap functions from github
#########################################################
library(formula.tools)

## Warning: package 'formula.tools' was built under R version 3.4.2

library(RCurl)

## Loading required package: bitops

s5 <- getURL("https://raw.githubusercontent.com/elegacy/lm.boot/master/wild.boot.R", ssl.verifypeer=FALSE)
source('http://elegacy14.weebly.com/uploads/1/0/2/7/102746882/bootstrapfunctions.r')
eval(parse(text=s5))
remove(s5)
set.seed(1414)
#########################################################

bootCrimeMod <- wild.boot(mergedData$Burglary.rate ~ mergedData$Larceny.theft.rate, B=10000)$bootEstParam
```

```r
print("Slope 95% CI")
```

```
## [1] "Slope 95% CI"
```

```r
BootCI(bootCrimeMod[,2], alpha=0.05) # 95% CI for slope coefficient
```

```
## [1] "(0.35, 0.37)"
```

```
## $lower
## [1] 0.3473637
##
## $upper
## [1] 0.3748735
```