Lab 4: Carrots, Sticks, and Crime

w203: Statistics for Data Science, Section 2

Krissy Gianforte & Dan Kent

18 December 2017

Introduction

G&K Associates have been retained by the [REDACTED] campaign to provide statistical modeling and analysis to understand the determinants of crime and generate policy suggestions that are applicable to local government. We, the principal investigators (K. Gianforte & D. Kent), utilized a pre-existing data set of crime statistics for a selection of counties.

The campaign has been considering two different approaches to reducing crime. The first, perhaps more obvious option is to reinforce deterrents. An increased police presence and lengthened prison sentences would eventually make crime simply unprofitable. However, this sort of increased force takes a toll on public opinion. The second approach, in contrast, would aim to decrease crime by facilitating better behaviors. Increased quality of life, prosperity, and happiness may keep people from reaching the desperation that fuels criminal acts.

This analysis explores each of the options (the "carrot" and the "stick").[1] We model crime as a function of "carrot" and "stick" indicators, and reveal the effects they are likely to have on crime rate.
NOTE: Without the ability to set up a true experiment, we cannot isolate these factors to determine the nature and direction of any patterns. That is, relationships seen through this analysis are not necessarily casual, and policy changes may not result in the desired effect despite our models' predictions.

Initial Exploratory Analysis

```
full_data <- read.csv("crime_v2_updated.csv", header = TRUE)
# Remove first column, since it is just an index. (The 'county' variable serves
# this purpose better, since every 'county' value is unique.)
full_data <- full_data[2:26]
```

We begin our Exploratory Data Analysis by inspecting all of the variables. We identify 25 different variables and confirm that these are the variables provided to us in the supplemental code book.

```
colnames(full_data)
```

```
##  [1] "county"   "year"     "crime"    "probarr"  "probsen"  "probconv"
##  [7] "avgsen"   "police"   "density"  "tax"      "west"     "central"
## [13] "urban"    "pctmin"   "wagecon"  "wagetuc"  "wagetrd"  "wagefir"
## [19] "wageser"  "wagemfg"  "wagefed"  "wagesta"  "wageloc"  "mix"
## [25] "ymale"
```

```
summary(full_data)
```

```
##      county           year         crime              probarr
##  Min.   :  1.0   Min.   :88   Min.   :0.005533   Min.   :0.1500
##  1st Qu.: 51.5   1st Qu.:88   1st Qu.:0.020604   1st Qu.:0.3642
```

[1] http://www.bocsar.nsw.gov.au/Documents/CJB/cjb54.pdf Omitted variables transcend a specific category and it is expected that variables could be both improved as well as added to our dataset.

```
##  Median :103.0   Median :88   Median :0.030002   Median :0.4222
##  Mean   :100.6   Mean   :88   Mean   :0.033510   Mean   :0.4106
##  3rd Qu.:150.5   3rd Qu.:88   3rd Qu.:0.040249   3rd Qu.:0.4576
##  Max.   :197.0   Max.   :88   Max.   :0.098966   Max.   :0.6000
##     probsen           probconv           avgsen           police
##  Min.   :0.09277   Min.   :0.06838   Min.   : 5.380   Min.   :0.0007459
##  1st Qu.:0.20495   1st Qu.:0.34422   1st Qu.: 7.375   1st Qu.:0.0012378
##  Median :0.27146   Median :0.45170   Median : 9.110   Median :0.0014897
##  Mean   :0.29524   Mean   :0.55086   Mean   : 9.689   Mean   :0.0017080
##  3rd Qu.:0.34487   3rd Qu.:0.58513   3rd Qu.:11.465   3rd Qu.:0.0018856
##  Max.   :1.09091   Max.   :2.12121   Max.   :20.700   Max.   :0.0090543
##     density             tax              west           central
##  Min.   :0.2034   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.5472   1st Qu.: 30.73   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.9792   Median : 34.92   Median :0.0000   Median :0.0000
##  Mean   :1.4379   Mean   : 38.16   Mean   :0.3778   Mean   :0.2333
##  3rd Qu.:1.5693   3rd Qu.: 41.01   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :8.8277   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##      urban             pctmin          wagecon         wagetuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.:10.024   1st Qu.:250.8   1st Qu.:374.3
##  Median :0.00000   Median :24.852   Median :281.2   Median :404.8
##  Mean   :0.08889   Mean   :25.713   Mean   :285.4   Mean   :410.9
##  3rd Qu.:0.00000   3rd Qu.:38.183   3rd Qu.:315.0   3rd Qu.:440.7
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##     wagetrd          wagefir          wageser          wagemfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.7   1st Qu.:285.6   1st Qu.: 229.3   1st Qu.:288.6
##  Median :203.0   Median :317.1   Median : 253.1   Median :321.1
##  Mean   :210.9   Mean   :321.6   Mean   : 275.3   Mean   :336.0
##  3rd Qu.:224.3   3rd Qu.:342.6   3rd Qu.: 277.6   3rd Qu.:359.9
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##     wagefed          wagesta          wageloc           mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:398.8   1st Qu.:329.3   1st Qu.:297.2   1st Qu.:0.08060
##  Median :448.9   Median :358.4   Median :307.6   Median :0.10095
##  Mean   :442.6   Mean   :357.7   Mean   :312.3   Mean   :0.12905
##  3rd Qu.:478.3   3rd Qu.:383.2   3rd Qu.:328.8   3rd Qu.:0.15206
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##      ymale
##  Min.   :0.06216
##  1st Qu.:0.07437
##  Median :0.07770
##  Mean   :0.08403
##  3rd Qu.:0.08352
##  Max.   :0.24871
```

From our high-level summary statistics and descriptions from the code book, we observe that some variables are categorical-ordinal (county, year), coded (west, central, urban), proportions or probabilities (probarr, probconv, probsen, pctmin, mix, ymale), averages (crime, avgsen, wagecon, wagetuc, wagetrd, wagefir, wageser, wagemfg, wagefed, wagesta, wageloc), and some are rates (crime, police, density, tax). However, all variables are represented as numeric data in the data frame; categorical and coded entries are represented as numeric 0's and 1's.

There are 90 unique counties represented in the data set, and each has values for all variables; no responses

are marked NA. It is possible, though, that the data set contains other values that represent non-applicable entries. As we introduce each variable into our models, we will inspect the values more carefully and address any such coding.

```
length(unique(full_data$county)) # number of unique counties
```

```
## [1] 90
```

```
nrow(full_data) # number of data rows
```

```
## [1] 90
```

```
any(is.na(full_data)) # any NA values?
```

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

The response/dependent variable of interest is *crime*: the quantity of "crimes committed per person." All values of this variable fall between 0 and 0.1, which aligns with our expectations; there do not appear to be any special coding conventions.

```
summary(full_data$crime)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020604 0.030002 0.033510 0.040249 0.098966
```

We exclude a few of the remaining 24 variables from our analysis, as they do not provide meaningful and reliable information:

- year - The year variable is ignored in our analysis as all the observations have the same value, 88, which the researchers understand as 1988 - perhaps the year of the data. As all of the values are the same across all observations, we ignore this variable.

- probarr, probconv, probsen - The variables involving probability are calculations in and of themselves; the code book describes these quantities as 'probability' values. The quotation marks in the definitions indicate that these are hypothetical, calculated values rather than truly observed values. For the purposes of this investigation, we wish to use measured, raw data. Therefore, these probability variables will be excluded from analysis.

- mix - the mix variable, described in the code book as "ratio of face to face/all other crimes" will be excluded as we believe this data to be associated with the response/dependent variable, *crime*, as opposed to a predictor/independent variable.

- county - The variable *county* is described as the "county identifier". This identification was used above to understand that each row of data represents a different county. Past that, though, this variable provides little value; the ordinal number associated with each observation does not provide any useful information about the associated data.

From this first pairing down of variables, we are left with the response variable *crime* and 19 independent variables.

```
data <- subset(full_data, select= -c(year, probarr, probconv, probsen, mix, county))
colnames(data)
```

```
##  [1] "crime"   "avgsen"  "police"  "density" "tax"     "west"    "central"
##  [8] "urban"   "pctmin"  "wagecon" "wagetuc" "wagetrd" "wagefir" "wageser"
## [15] "wagemfg" "wagefed" "wagesta" "wageloc" "ymale"
```

We have grouped these variables into categories based upon their type of effect in our models: Authority, Demographics, Geography, and Economics. Moving forward, we will discuss variables based upon these groupings.

Authority

Government-controlled deterrents to crime are included in the "Authority" category. This includes concepts meant to prevent crime from occurring (such as police presence) as well as things that increase the penalty to crime (such as prison sentence length). From the code book, we identify two variables in this category:

1. *police* - police per capita
2. *avgsen* - average sentence, in days

Demographics

The "Demographics" category includes multiple variables that describe the cultural and personal characteristics of the county's population. In general, these variables are expressed as proportions of the population. From the code book, we identify two variables in this category:

1. *ymale* - proportion of county males between the ages of 15 and 24
2. *pctmin* - proportion that is minority or non-white

Geography

The "Geography" category includes all variables that describe the land, location, and housing of a county. There are a number of such variables in this data set, and their expression varies from normalized values (such as *density*, expressed in people per sq mile) to binary indicators (such as the *urban*, *west*, and *central* indicators). The geographic variables are:

1. *density* - people per sq. mile
2. *west* - indicator, =1 if in western part of the state
3. *central* - indicator, =1 if in central part of the state
4. *urban* - indicator, =1 if in Standard Metropolitan Statistical Area

Note that the *urban* indicator is not mutually exclusive with the *west* and *central* location indicators. Some counties are labelled as both *urban* and *west*, for example. However, the *west* and *central* labels are mutually exclusive; a county is included in only one of the location categories.

```
# Urban does not exclude west/central labels
data[data$urban == 1 & data$west == 1, c("urban", "west")]
```

```
##    urban west
## 29     1    1
## 31     1    1
## 36     1    1
## 53     1    1
## 83     1    1
```

```
# West & Central labels are mutually exclusive
data[data$west == 1 & data$central == 1, c("west", "central")]
```

```
## [1] west    central
## <0 rows> (or 0-length row.names)
```

Without knowing the specific state in question here, it is difficult to understand the meaning of the *west* and *central* geographical impact. However, distance from a major city is a well-known and universally understandable metric. Therefore, our analysis focuses on the *urban* indicator variable.

Economics

Finally, wages and taxes are included in the "Economics" category. In our analysis, these variables are meant to capture the prosperity of a county, and provide an operationalization for quality-of-life. The specific variables are:
1. *wagecon* - weekly wage, construction
2. *wagetuc* - weekly wage, transportation, utilities, communications
3. *wagetrd* - weekly wage, retail trace
4. *wagefir* - weekly wage, finance, insurance & real estate
5. *wageser* - weekly wage, service industry
6. *wagemfg* - weekly wage, manufacturing
7. *wagefed* - weekly wage, federal employees
8. *wagesta* - weekly wage, state employees
9. *wageloc* - weekly wage, local government employees
10. *tax* - tax revenue per capita

There are some concerning limitations in understanding the parameters of the *tax* variable. Specifically, we are concerned that without further information, the tax variable might co-mingle different types of tax data, representative of not only personal taxes, but also taxes of businesses. The wage variables have some uncertainty as well: because we don't know the population of the counties or denominator of the average wages, the "per capita/averages" could be significantly skewed or not reliably comparable across all counties. Forced to choose one economic metric, we selected the wage variables for our models. We also created several custom wage variables, to capture a slightly broader picture of economic disparity between citizens. In particular, we created:

- *wage_avg* - the average wage across all professions
- *wage_private* - the average wage in the private domain (wage variables 1-6)
- *wage_public* - the average wage of government employees (wage variables 7-9)

Note that without raw data we cannot weight these average calculations properly, so our 'average of averages' values are surely inaccurate. Still, the general concept is valuable. For this analysis, we knowingly sacrifice a bit of accuracy in order to explore economic effects. Should those effects prove to be significant, we would desire to conduct further study and collect the exact wage data of interest.

Model Building

This analysis presents three different models to describe how crime is affected by various factors. The first model is quite simple, and uses just two variables to operationalize positive ("carrot") and negative ("stick") control of crime. This model is an over-simplification, but such a simple picture often proves useful in presenting ideas to large groups of campaign supporters and investors.

The second model is a more accurate depiction of the complex ecosystem around crime. It incorporates related and entangled variables where relevant, and thereby improves on the predictive capabilities of the first model. It also controls for demographic and geographic factors.

Finally, we present a third model with all related factors included. This model is provided largely as a baseline, and helps to demonstrate the usefulness of our second proposed model.

For each model, we assess the Multiple Linear Regression assumptions MLR.1-6 to ensure validity. In all cases, we are able to pass the first assumption, MLR.1, since we define the models into a linear form. The veracity of the second assumption, MLR.2, i.e. random sampling, IID, is difficult to discern based upon the data and additional information we have at our disposal. As we did not conduct the data collection and further have limited information about the variables, we proceed cautiously in this analysis with the assumption that our data meets MLR.2.

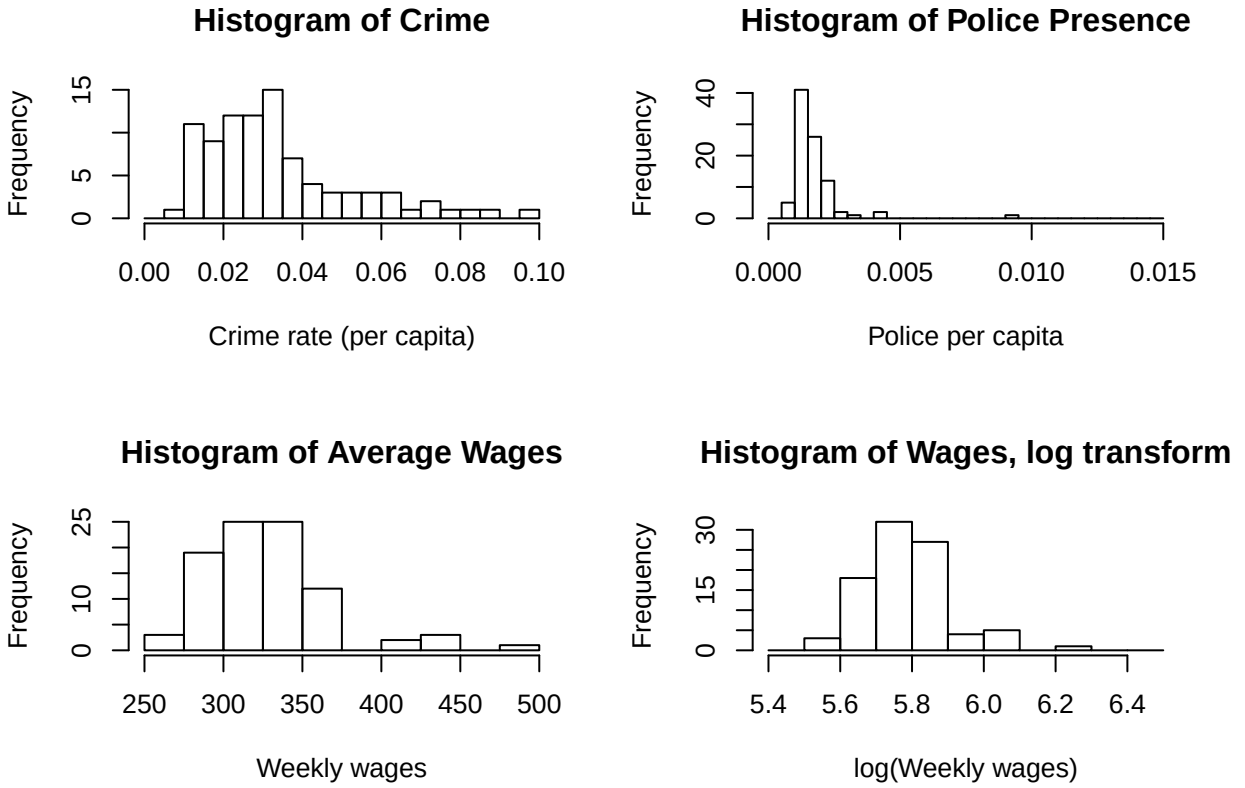
Model 1:

Specification

Crime is modeled here as a function of only two variables: police presence (a deterrent) and average wages (to represent positive incentives).

Histograms of each of these variables allow us to identify skew and make any required transformations. NOTE: It is quite common to transform salary/wage variables using a logarithm. In anticipation of that, a histogram of *log(wage_avg)* is provided proactively here.

```
# Create the variable for average wage
wagevars = c("wagefed", "wagesta", "wageloc", "wagecon", "wagetuc", "wagetrd", "wagefir",
             "wageser", "wagemfg")
data$wage_avg = apply(data[,wagevars], 1, mean)

#Histogram of each variable in the model
par(mfrow=c(2,2))

hist(data$crime, main = "Histogram of Crime", xlab = "Crime rate (per capita)",
     breaks = seq(0, 0.10, 0.005))

hist(data$police, main = "Histogram of Police Presence", xlab = "Police per capita",
     breaks = seq(0,0.015, 0.0005))

hist(data$wage_avg, main = "Histogram of Average Wages", xlab = "Weekly wages",
     breaks = seq(250, 500, 25))

# Create a log(wage) variable
data$log_wage_avg <- log(data$wage_avg)
hist(data$log_wage_avg, main = "Histogram of Wages, log transform",
     xlab = "log(Weekly wages)", breaks = seq(5.4, 6.5, 0.1))
```



The dependent variable of interest, *crime*, has a somewhat skewed distribution, with many counties on the lower end of the spectrum and a tail extending into higher rates. Transforming this variable using a logarithm would likely result in more normal data. The transformed log(*crime*) would still be understandable; a change in log(*crime*) would simply represent a percent change. With that in mind, we transform the output variable and create a new *log_crime* variable.

```
# Create log crime variable
data$log_crime = log(data$crime)

hist(data$log_crime, main = "Histogram of Crime, log transform",
     xlab = "log(Crime rate(per capita))")
```

The *police* variable is approximately normally distributed, except for a single value near 0.010 (1 police officer for every 10 citizens). We will incorporate the variable into the model without a transformation, given its nearly-normal distribution.

As expected, the histogram of log(*wage_avg*) appears more normal than the untransformed plot; a logarithm helps reduce the impact of high-earners. A log transform of pay also helps with intuitive interpretation of the model (an X% increase in wages is quite easy to understand). With both of those advantages, we select the log(*wage_avg*) variable for the model.

Our model takes the following form:

$$log(Crime) = \beta_0 + \beta_1 \cdot Police + \beta_2 \cdot log(AvgWages)$$

```
Model1 <- lm(log_crime ~ police + log_wage_avg, data = data)
```

Assumptions

We proceed with testing assumptions to understand the robustness of our created model.

First we identify if the data we have available fits our model theory. As our model uses a "carrot" and "stick" approach we investigate the former and later variables. We identify that police presence ("stick") is appropriate for our model, but our "carrot" is potentially of concern as we are taking the "average of averages" in this context. (See more detailed discussion in the the "Economics" section above.) We are presented with no other data and consequently will proceed with the understanding that this may limit the full interpretability of our model.

The first two Multiple Linear Regression assumptions, MLR.1-2, are discussed above for all models. Here we begin with the the third assumption, MLR.3, the assumption of no perfect colinearity. As we have multiple variables, we will look at the R^2 value, tolerance, and VIF.

```
summary(Model1)$r.squared
```

```
## [1] 0.1166539
```

Our R^2 value is not approaching 1 therefore there is no threat of multicolinearity evidence there.

```
(tolerance1 <- 1-summary(Model1)$r.squared)
```

```
## [1] 0.8833461
```

Our tolerance value is also not below 0.1, which therefore does not indicate any serious multicolinearity. We now calculate our VIF:

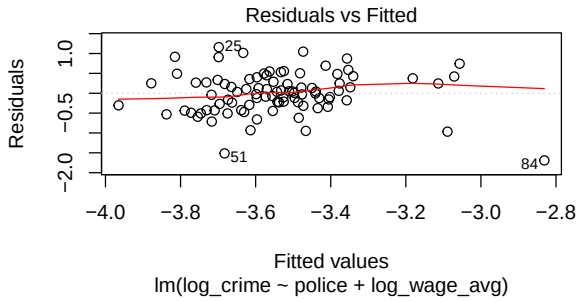
```
(VIF1 <- 1/tolerance1)
```

```
## [1] 1.132059
```

Finally, our VIF is not above four, let alone 10, indicating that we do not have any issues of multicolinearity.

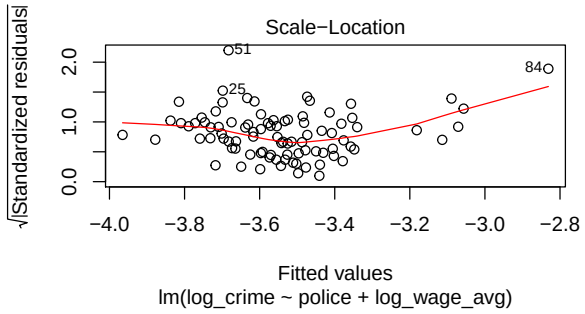
The next assumption we examine is that of the Zero Conditional Mean, MLR.4. We first identify if there are any omitted variables - as we have a limited data set with respect to different variables, we are caught in a bit of a bind. Theoretically there are many more "carrot" and "stick" variables that we could use in our model, however some of these are not available to us and we want to reserve others for our second model. We will proceed investigating MLR.4 with caution based on our conclusions here.

We proceed to create a residuals versus fitted values plot:



From our plot, we observe a spline curve that is fairly flat. We notice some curvature towards the positive side of the plot, but observe that this could be explained by the limited number of data points in this region. We consequently will assume Zero Conditional Mean and continue with the analysis.

Our next tested assumption is MLR.5, the assumption of homoskedacity. From the above residuals versus fitted values plot, we can observe that the band has a roughly uniform thickness, with the potential exception of two points in the lower area of the plot. We will proceed cautiously with further analysis of homoskedasticity but also use robust standard errors (see infra.)



Our scale-location plot also depicts a relatively horizontal band of points, except, as seen in the residuals vs fitted values plot, for a handful of data points in the far right region of the plot. Of note, data points 51 and 84 appear to be the most egregious aberrations. With this in mind, we continue our analysis cautiously and return to analyze points 51 and 84 later.

```
bptest(Model1)
```

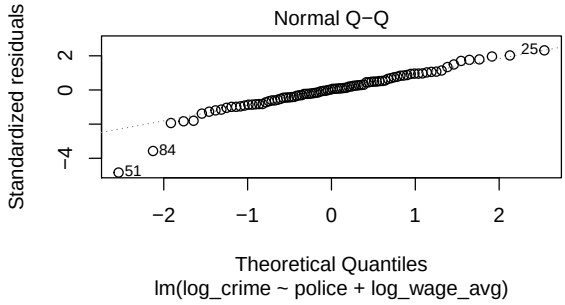
```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  Model1
## BP = 19.36, df = 2, p-value = 6.253e-05
```

From our Breusch-Pagan test, we must reject the null hypothesis of homoskedasticity as we have a p-value of less than .05, indicating that there is evidence supporting heteroskedasticity. With that in mind, we will use heteroskedasticity-robust methods of analysis going forward. (See the "Summary of Models" section.)

We proceed with the analysis of MLR.6, the normality of errors, by investigating the residual plot, which is roughly normally distributed:

Our QQ test shows that the majority of points fall on the diagonal line with some departure at the upper and lower extremes. Again, points 51 and 84 are the most obvious exceptions to the fit.

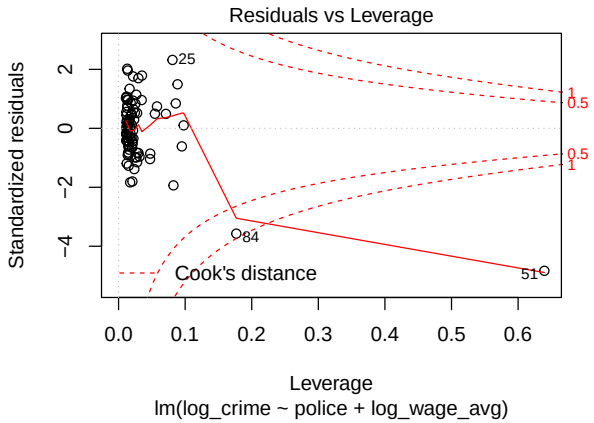


Since our sample is approaching 100 data points, we can rely on asymptotic properties of OLS, including the central limit theorem, to help with our testing of MLR.6.

```
length(data$crime)
```

```
## [1] 90
```

Finally, we use a Residuals vs Leverage plot to search for points with disproportionate leverage. We see two such points: again points 51 and 84.



We investigate these two points, to see whether a coding issue could be responsible for their effect on the model.

```
data[c(51, 84), c("crime", "police", "wage_avg")]
```

```
##        crime     police wage_avg
## 51 0.0055332 0.00905433 347.7498
## 84 0.0108703 0.00122210 495.9648
```

```
head(sort(data$crime))
```

```
## [1] 0.0055332 0.0106232 0.0108703 0.0119154 0.0121033 0.0126662
```

```
head(sort(data$wage_avg, decreasing = TRUE))
```

```
## [1] 495.9648 441.6914 439.1508 425.8236 422.9731 411.2707
```

Point 51 has the absolute lowest crime rate in our data set: 0.00553. The next lowest value of *crime* is 0.0106 - nearly twice the value of point 51. The low crime rate *does* seem to be a plausible value, even if it is extreme. Without further knowledge of county 51, we cannot remove point 51 as a true outlier or coding error. We must retain it in the model.
Point 84 has the third lowest crime rate in the data set: 0.0109. Interestingly, county 84 has the absolute highest average wages: 495.96. This again is a plausible value, if extreme. We leave point 84 in the model.

These two data points do create some concerns around leverage and outliers, However, as our sample approaches 100 data points, as mentioned above, we can rely on asymptotic properties of OLS including the central limit theorem.

Having satisfied the six Multiple Linear Regression model assumptions, we expect the Model 1 coefficients to be the best linear unbiased estimators (BLUE). That is, the coefficient values are accurate in expectation ($E(\beta_j) = \beta_j$) and have small variance as compared to other linear unbiased estimators. In fact, our large sample size (n>30) and asymptotic properties allow us to make this claim based only on the first four assumptions.
* (MLR.1) linear model specification
* (MLR.2) random sampling
* (MLR.3) no perfect multicollinearity
* (MLR.4) zero conditional mean
* (MLR.5) homoskedasticity - *addressed by heteroskedasticity-robust methods*
* (MLR.6) normality of residuals

Model 2:

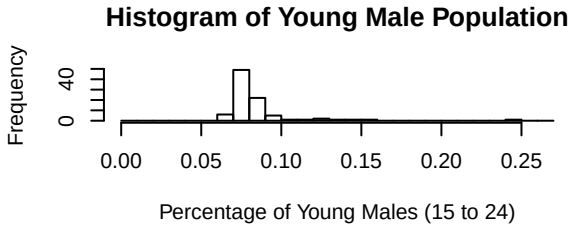
Specification

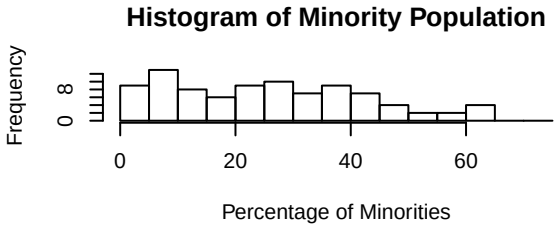
This model adds a bit of complexity to the first, still within the "carrot and stick" paradigm, in order to better capture the full range of variables that affect crime. We incorporate interaction terms and covariates where they might affect the model. In particular, we include at least one variable from each category discussed supra, in order to capture other factors in society and control for them:

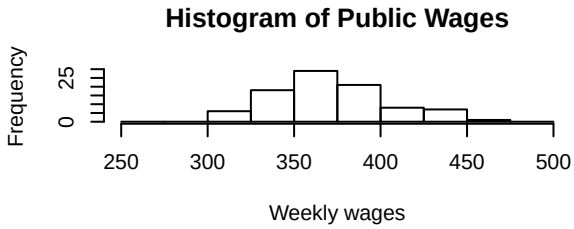
- Demographics: operationalized by the number of young males in the county, *ymale*, and the percentage of minority citizens, *pctmin*
- Geography: operationalized by population density, *density*
- Economics: in place of the incentive variable *wage_avg* captured in Model 1, the average wages of government and non-government employees, *wage_public* and *wage_private* respectively, are put into the model as separate terms. This allows us draw out and quantify the prosperity of an average citizen in the private domain (via *wage_private*).

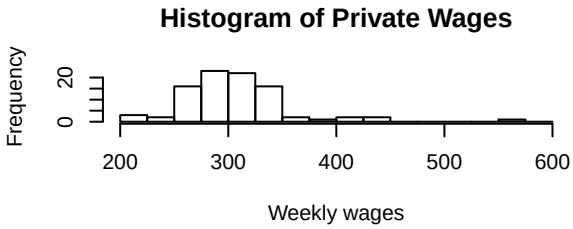
Histograms of each of these variables allow us to identify skew and make any required transformations.

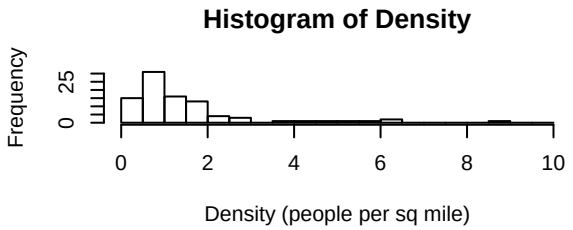
```
# Create the variables for public & private wages
pubwages = c("wagefed", "wagesta", "wageloc")
privwages = c("wagecon", "wagetuc", "wagetrd", "wagefir", "wageser", "wagemfg")
allwages = c(pubwages, privwages)
data$wage_public = apply(data[,pubwages], 1, mean)
data$wage_private = apply(data[,privwages], 1, mean)

data$wage_max = apply(data[,allwages], 1, max)
data$wage_min = apply(data[,allwages], 1, min)
data$wage_range <- (data$wage_max - data$wage_min)

#Histogram of each NEW variable
par(mfrow=c(3,2))

hist(data$ymale, main = "Histogram of Young Male Population",
     xlab = "Percentage of Young Males (15 to 24)", breaks = seq(0, 0.27, 0.01))

hist(data$pctmin, main = "Histogram of Minority Population",
     xlab = "Percentage of Minorities", breaks = seq(0,75,5))

hist(data$wage_public, main = "Histogram of Public Wages",
     xlab = "Weekly wages", breaks = seq(250, 500, 25))

hist(data$wage_private, main = "Histogram of Private Wages",
     xlab = "Weekly wages", breaks = seq(200, 600, 25))

hist(data$density, main = "Histogram of Density",
     xlab = "Density (people per sq mile)", breaks = seq(0,10,0.5))
```











The *ymale* and *density* variables have nearly normal distributions with a few values sitting far down the right tail. *Density* in particular seems to have a cutoff value at zero; the distribution would perhaps appear more normal if it were allowed to extend leftward. Of course density cannot go below zero, so the cutoff is appropriate.

The *pctmin* variable has a somewhat uniform distribution between 0 and 50%.

The distribution of public wages, *wage_public*, is quite normal, so we will include it in the model as-is. For the sake of comparability, we will also include the *wage_private* variable as-is. This keeps both in terms of dollar change (as opposed to percentage changes).

Our model takes the following form:

$$\begin{aligned} log(Crime) = \beta_0 + \beta_1 \cdot Police + \beta_2 \cdot Ymale + \beta_3 \cdot PctMinority \\ + \beta_4 \cdot Density + \beta_5 \cdot PublicWages + \beta_6 \cdot PrivateWages \end{aligned}$$

```
Model2 <- lm(log_crime ~ police + ymale + pctmin + density + wage_public + wage_private,
             data = data)
```

Assumptions

Following the same assumption test protocol we conducted in the above section for Model 1, we investigate any violations of our assumptions.

```
summary(Model2)$r.squared
```

```
## [1] 0.5526714
```

```
(tolerance2 <- 1-summary(Model2)$r.squared)
```

```
## [1] 0.4473286
```

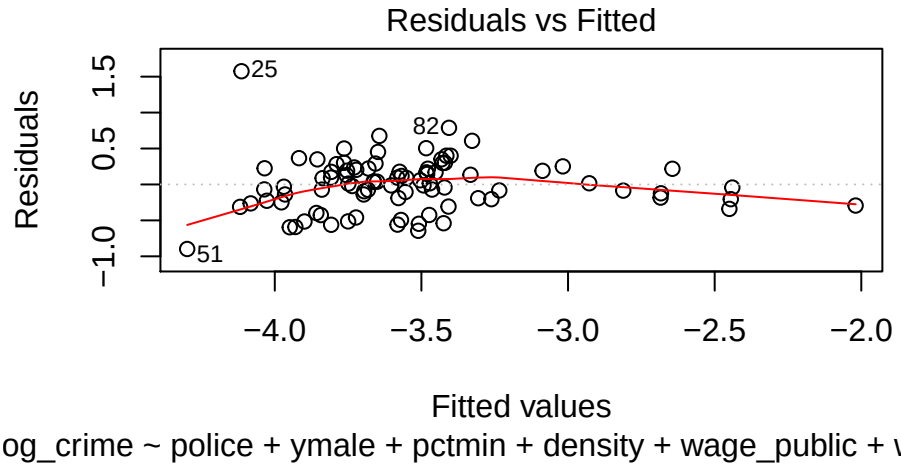
```
(VIF2 <- 1/tolerance2)
```

```
## [1] 2.235493
```

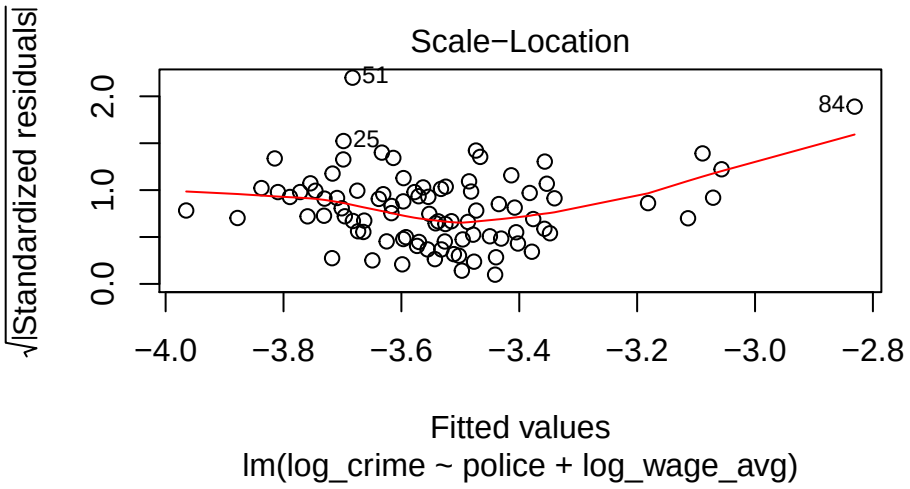
```
bptest(Model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Model2
## BP = 18.886, df = 6, p-value = 0.00436
```

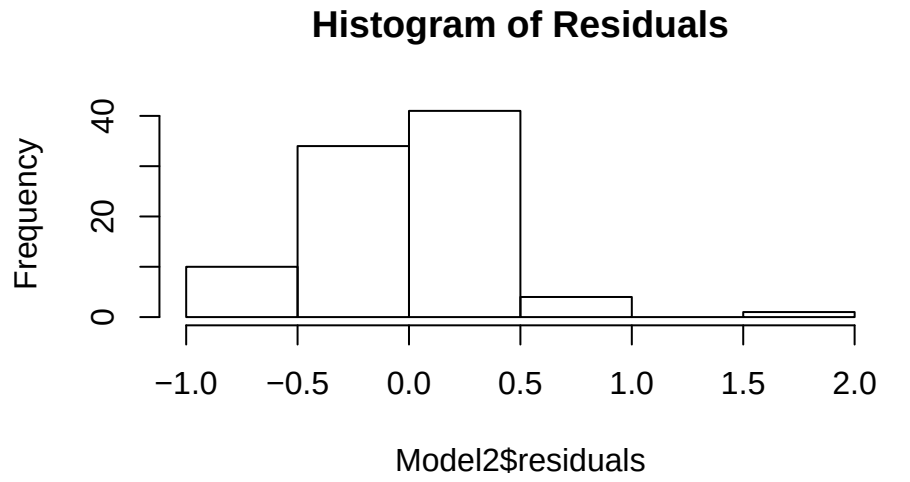
```
#Diagnostic Plots

# Residuals vs Fitted Values
plot(Model2, 1)
```



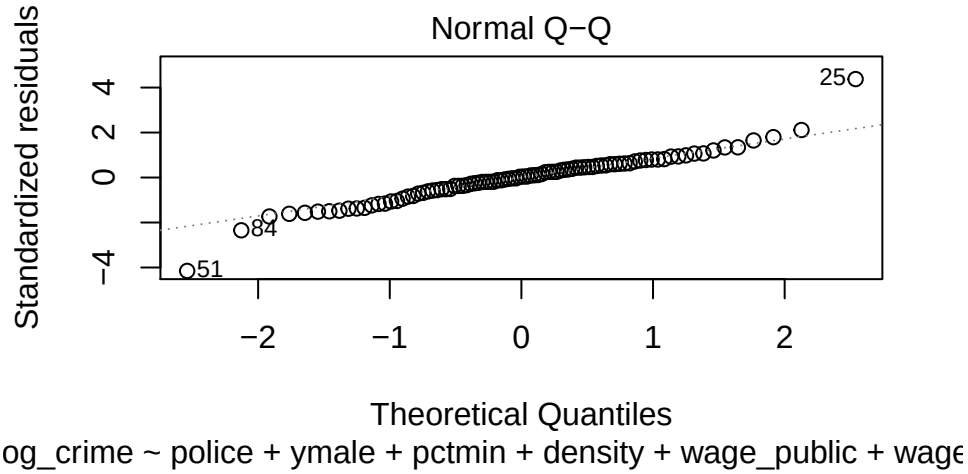

```
# Scale-Location
plot(Model1, 3)
```



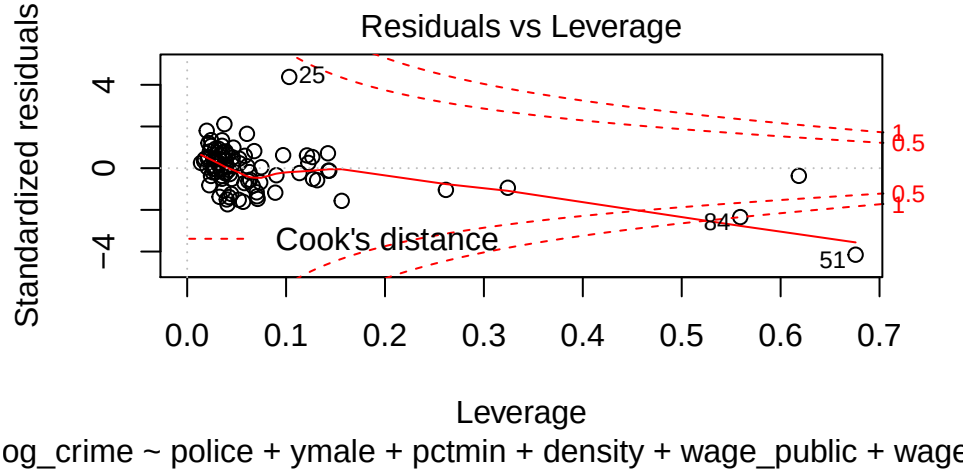

```r
# Histogram of residuals
hist(Model2$residuals, main = "Histogram of Residuals")
```




```r
# Normal Q-Q
plot(Model2, 2)
```




```
# Residuals vs Leverage
plot(Model2, 5)
```



From our data outputs, we identify some heteroskedasticity, as we observed above and consequently will switch to robust standard errors (see infra). In our QQ plot analysis, we note general conformity to the diagonal line, with the exception of points 51, 84, and 25. Moving further in our analysis, we again identify that points 51 and 84, while legitimate data, exert high leverage and influence.

As with Model 1, we need only satisfy the first four MLR assumptions to know that the OLS estimated coefficients are unbiased. Still, we reckon that we are able to cautiously accept *all* assumptions for Model 2 (using Heteroskedasticity-robust methods to address MLR.5.

Model 3:

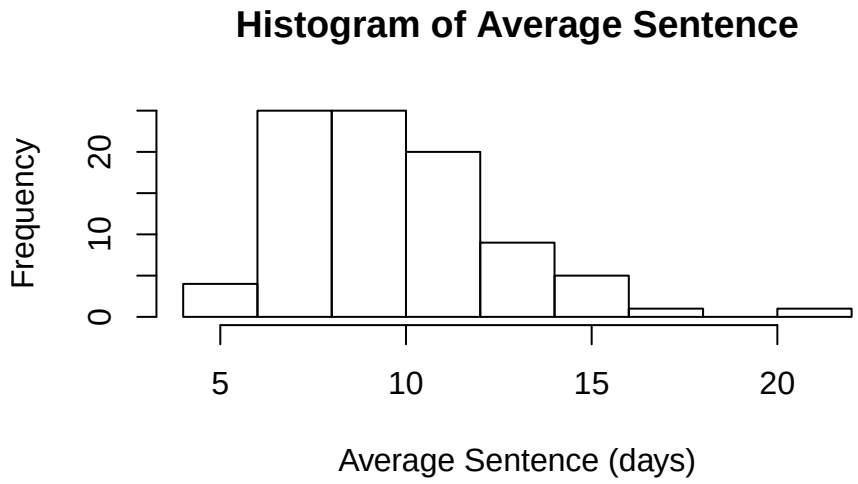
Specification

This model expands upon the operationalization of deterrents to crime by adding the *avgsen* variable. Note

that we did not add this variable into earlier models because of a lack of confidence in its derivation. (Does it represent sentence length averaged over convictions? Or over all citizens - even those with no crimes? Is it normalized for the severity of the crime, or are sentences served for murder and public intoxication weighted equally?)
Additionally, this model incorporates the *urban* indicator variable as an interaction term with the *ymale* variable, to differentiate the activities of young males in different areas.

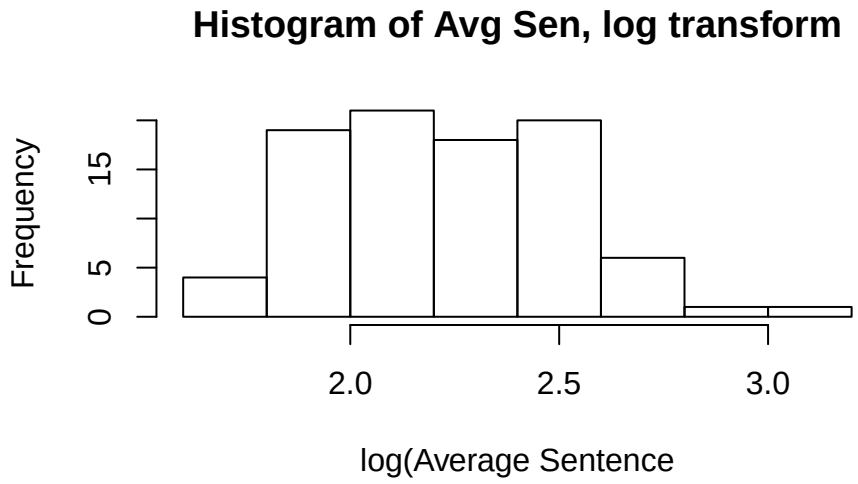
A histogram of the *avgsen* variable allows us to identify skew and make any required transformations.

```
#Histogram of average sentence & percent minority
hist(data$avgsen, main = "Histogram of Average Sentence", xlab = "Average Sentence (days)")
```



We proceed in transforming this variable, due to its skew.

```
#Histogram of average sentence, log transform
hist(log(data$avgsen), main = "Histogram of Avg Sen, log transform",
     xlab = "log(Average Sentence")
```



Our model takes the following form:

$$log(Crime) = \beta_0 + \beta_1 \cdot Police + \beta_2 \cdot Ymale + \beta_3 \cdot PctMinority + \beta_4 \cdot Density + \beta_5 \cdot PublicWages + \beta_6 \cdot PrivateWages + \beta_7 \cdot Urban + \beta_8 \cdot Urban * ymale + \beta_9 \cdot log(AvgSentence)$$

```
Model3 <- lm(log_crime ~ police + ymale + pctmin + density +
               wage_public + wage_private + urban + urban*ymale +
               log(avgsen), data = data)
```

Assumptions

```
summary(Model3)$r.squared
```

```
## [1] 0.563588
```

```
(tolerance3 <- 1-summary(Model3)$r.squared)
```

```
## [1] 0.436412
```

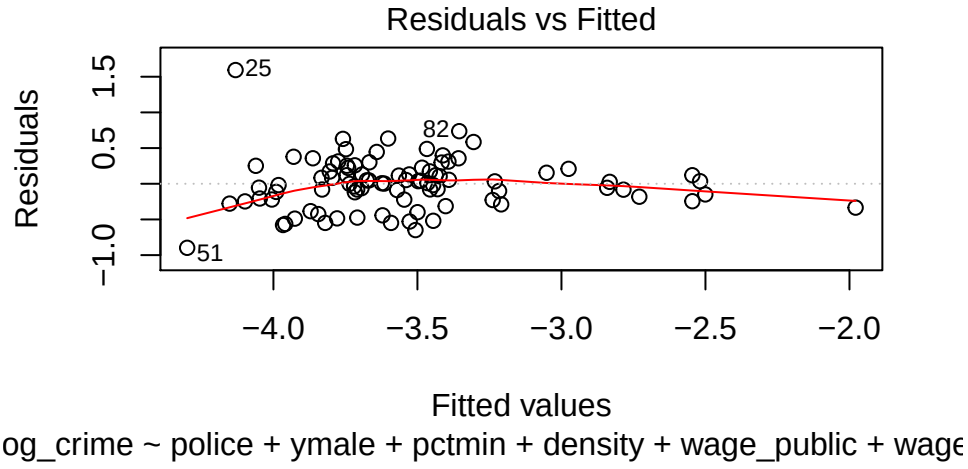
```
(VIF3 <- 1/tolerance3)
```

```
## [1] 2.291413
```

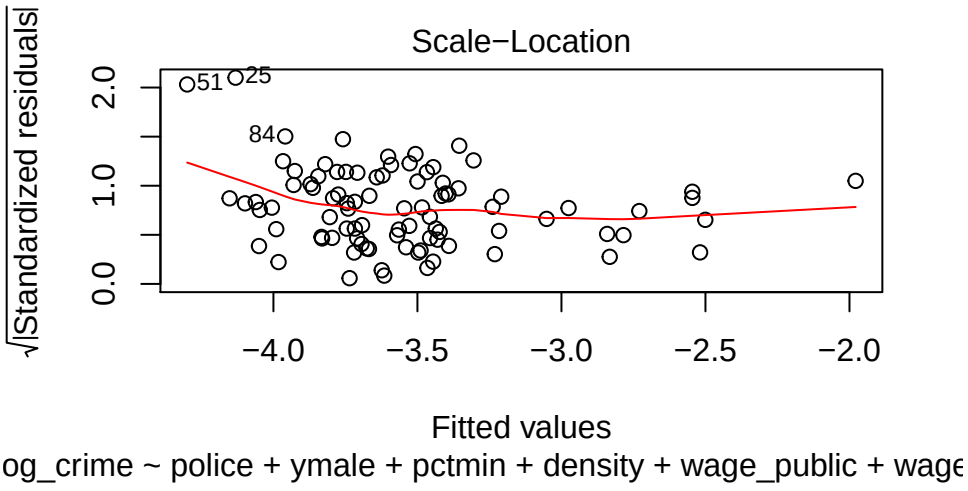
```
bptest(Model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Model3
## BP = 19.74, df = 9, p-value = 0.01958
```

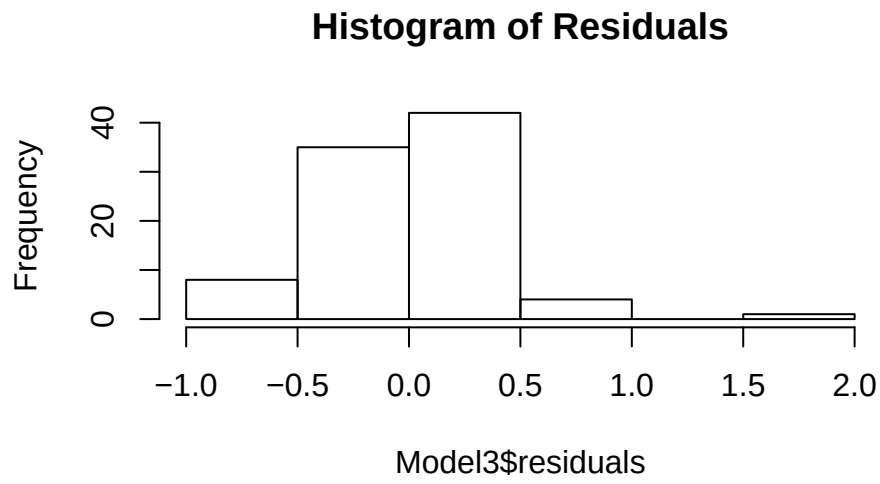
```
#Diagnostic Plots

# Residuals vs Fitted Values
plot(Model3, 1)
```



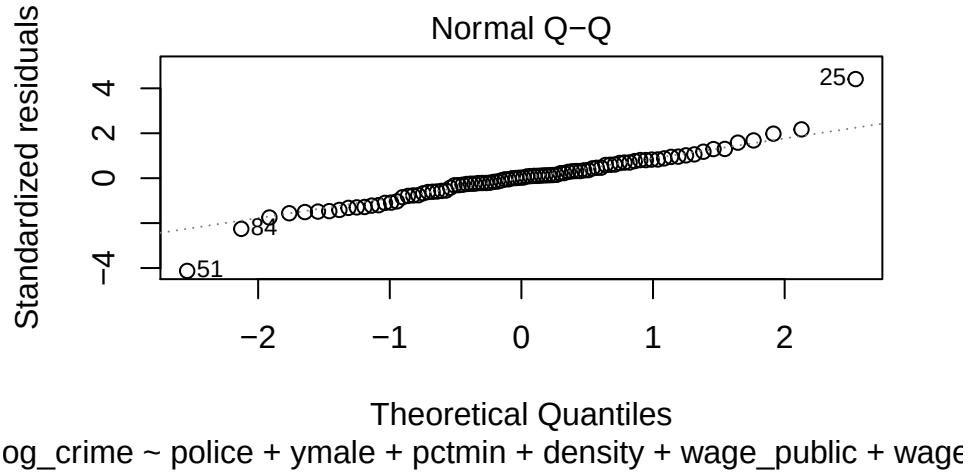

```
# Scale-Location
plot(Model3, 3)
```



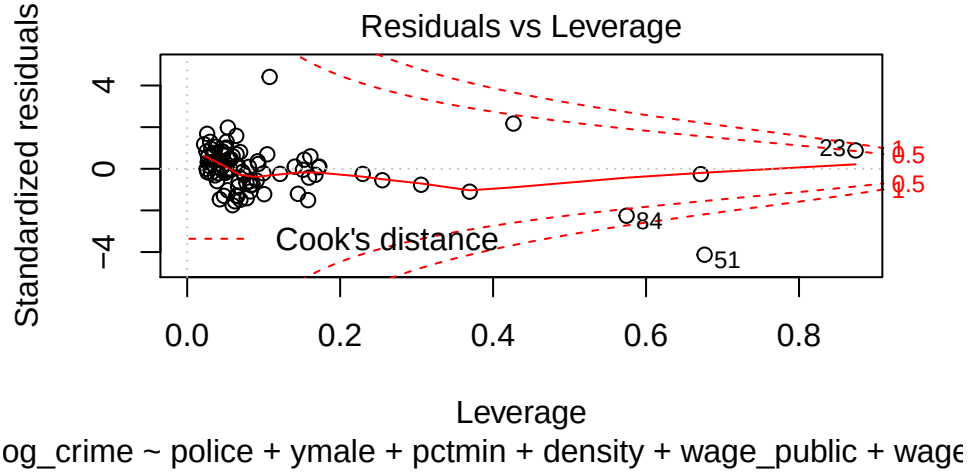

```
# Histogram of residuals
hist(Model3$residuals, main = "Histogram of Residuals")
```




```
# Normal Q-Q
plot(Model3, 2)
```




```
# Residuals vs Leverage
plot(Model3, 5)
```



From our data outputs from Model 3, we identify some heteroskedasticity, as we observed above in Models 1 and 2 and similarly switch to robust standard errors (see infra). Our scale-location plot calls attention to points 51, 25, and potentially 84. In our QQ plot analysis, we note general conformity to the diagonal line, with the exception of points 51, 84, and 25, as observed in Model 2. We identify that points 51 and 84, while legitimate data, exert high leverage and influence, as does point 23, a new finding for Model 3.

The first four MLR assumptions are again satisfied for Model 3, so we know that the OLS estimated coefficients are unbiased and efficient. However, the presence of high-leverage and high-influence points 51, 84, and 25 is concerning. Model 3 is likely not the best choice of specification; the regression table below will help assess further.

Summary of models

Regression Table

```
# Use heteroskedasticity-robust methods
se.model1 = sqrt(diag(vcovHC(Model1)))
se.model2 = sqrt(diag(vcovHC(Model2)))
se.model3 = sqrt(diag(vcovHC(Model3)))

# Stargazer output
stargazer(Model1, Model2, Model3, type = "latex", header = FALSE,
          title = "Linear Models Predicting Crime",
          se = list(se.model1, se.model2, se.model3),
          star.cutoffs = c(0.5, 0.01, 0.001),
          omit.stat = "f")
```

```
c(AIC(Model1), AIC(Model2), AIC(Model3))
```

[1] 143.21778 89.97965 93.75606

Significance

Statistical Significance

From an analysis of our three models, we observe an increase in R^2 with each successive model. However, the R^2 increase in Model 3 over Model 2 is not actually an improvement, as indicated by the Adjusted R^2 and AIC comparison. (AIC values are 143.22, 89.98, and 93.76 respectively; a lower AIC indicates a better model.) This informs us that Model 2 offers the best balance between parsimony and best-fit.

In our first model, we observe the only variable that has statistical significance at the .05 level and demonstrates that an effect is present is that of *log_wage_avg*, illustrated by a very low p-value. In other words, our other variable, *police* does not present any statistically significant effect on our response variable, *crime*.

In our second model, we see that a number of our variables are statistically significant at the .05 level - *ymale*, *pctmin*, *density*, and *wage_public*, demonstrating that there is an effect present on our response variable, *crime*, with these variables.

Our final model, Model 3, depicts a similar picture to our second model, where there are the same four variables that are statistically significant at the .05 level - *ymale*, *pctmin*, *density*, and *wage_public*, demonstrating again that there is an effect present on our response variable, crime, with these four variables.

Practical Significance

In our first model, we observe some practical significance with the *log_wage_avg*: the coefficient is 1.653. This value is too large to interpret using the "percent change" understanding of logarithmic specification (this works for changes less than about 20%).

Our second model depicts that every unit increase in the proportion of young adult males in a county is reflected by a 4.672 times increase in log of crimes per person. This is deduced from the *ymale* variable. *Density* also is associated with an increase in crimes; for every additional person per square mile, we observe a 21.2% increase in crimes committed per person. (This coefficient is small enough to use the percentage change interpretation of the logarithm.) The other statistically significant variables, *pctmin* and *wage_public*, have an effect of .009 and .003, respectively, so there is little practical significance, even though these data are statistically significant.

Table 1: Linear Models Predicting Crime

	Dependent variable:		
	log_crime		
	(1)	(2)	(3)
police	−33.852	−41.626	−34.094
	(366.767)	(248.569)	(249.810)
log_wage_avg	1.653*		
	(1.006)		
ymale		4.672**	4.243**
		(1.520)	(1.532)
pctmin		0.009**	0.009**
		(0.003)	(0.003)
density		0.212***	0.264**
		(0.057)	(0.097)
wage_public		0.003*	0.003*
		(0.002)	(0.002)
wage_private		−0.001	−0.002
		(0.002)	(0.002)
urban			−0.816
			(1.580)
log(avgsen)			−0.035
			(0.149)
ymale:urban			5.528
			(15.774)
Constant	−13.047*	−5.188***	−5.133***
	(5.545)	(0.685)	(0.707)
Observations	90	90	90
R^2	0.117	0.553	0.564
Adjusted R^2	0.096	0.520	0.514
Residual Std. Error	0.522 (df = 87)	0.380 (df = 83)	0.382 (df = 80)
Note:			*p<0.5; **p<0.01; ***p<0.001

Finally, in Model 3, we observe nearly identical outcomes as our second model, stated supra. This model depicts that for every unit increase in the percentage of young adult males in a county, the log of crimes increases by 4.243, a reduction in effect from our previous model. This reduction is potentially due to the inclusion of the *urban* x *ymale* interaction term; some of the effect of young males may be accounted for by the effects of a city.

The practical significance of *density* increases slightly - each additional person per square mile is associated with a 26.4% increase in crimes committed per person. As stated in Model 2, the practical significance of the other statistically significant variables, *pctmin* and *wage_public*, is very small.

While not statistically significant, across all three models police presence was associated with a strong reduction in our response variable, crimes committed per person. The coefficient of the *police* term was -33.852, -41.625, and -34.094 across Models 1, 2, and 3, respectively. While we cannot claim statistically significant relationships, due to the size of the association, we believe further research is warranted.

Causality & Omitted Variables

The data available to us presented significant barriers with respect to identifying causality as our models were limited solely to the supplied data set. This resulted in us drawing conclusions and building models based on data collected by a third party, who's veracity and trustworthiness we cannot confirm, let alone of that of the data. Further, we do not know if there were additional confounding elements to their collection of the data and whether that had an impact on the responses provided. What we can say is that there is a high degree of likelihood that there are other variables, omitted in the data set and consequently our modeling, that could have an impact on our response variable *crime*, the number of crimes committed by person.

Our thesis was to test our model developed around that of the "carrot and stick," based on the data we collected. From our outcomes, our model fell short of our expectations. We believe that it is in the spirit of statistics and research that further modeling and hypothesizing should take place in the construct of such model dialogue. For example, one study [2] identified a number of distal and proximate causes to create a number of models, such as that of the influence of parent-child attachment and delinquent peer model, or the economic stress, influence of parent-child attachment, and delinquent peer model. We feel that this line of inquiry should continue.

For example, in our Economics category, the weekly wages of nine different industries were provided. However, as good students of statistics, we know that this sort of variable - which we discern as average weekly wages, is a poor statistic, as it is one that is frequently skewed by both high and low earners. Instead of this "weekly wage" per industry, perhaps better variables would rather be the percentage of population in each tax bracket - which would show a distribution of income across the county's population. We would expect that this more accurate measurement would be associated with a slightly positive direction, as those more affluent areas would be the target of more crime, but because of additional resources of those areas, crime would be suppressed to a degree through security activity. Alternatively, it is popular knowledge that crime is sometimes associated with lower-levels of income. A different variable that might provide for better modeling is the percentage of the population that is on public welfare assistance programs or the percentage of population who are classified as living at or below the federal poverty line. We hypothesize that this would be positively correlated with crime. An additional variable that could potentially predict crime could be the "street price" of certain drugs and other illegal substances, also positively correlated with our response variable. Unemployment and income inequality are also both variables that could represent economic conditions within a county and help us build a model that better predicts *crime*, both also positively correlated with our dependent variable.

Further details about our populations in each county could provide additional variables that help us understand and develop a model to predict *crime*. For example, new variables, omitted currently could include percentage of population in a gang, percent of population previous felons, percent of population with

[2] http://www.bocsar.nsw.gov.au/Documents/CJB/cjb54.pdf Omitted variables transcend a specific category and it is expected that variables could be both improved as well as added to our dataset.

previous misdemeanors, percent population on parole, and percentage of population who engage in substance abuse, all variables which we believe would contribute to an increase in crime in the observed areas. Further demographic variables that could aid our model include high-school drop-out rates, female-to-male ratio, percentage of single parents, and percentage of single female parents, all which are hypothesized to positively impact *crime*.

With respect to our Authority category, we were surprised at the lack of statistical significance of the effect of *police*, the number of police per capita. While we were able to observe a significant practical effect (even though it was not statistically significant) further inquiry and investigation is warranted. Perhaps a more effective variable would be that not of the quantity of general "police," but rather the quantity of officers exclusive of those who are involved in tasks that include "back-office" work, i.e. those police officers who canvas specific neighborhoods with on a particular beat. Other omitted variables that could play a significant role in the outcome of our response variable could be the budget of the police department, as police departments with larger budgets could be more efficient than those with large populations of police officers, but with few resources.

As part of the role of this endeavor is to identify policy and program activities to reduce crime, additional variables, particularly those which might be categorized under Authority, should be investigated. For example, specific technologies implemented such as predictive policing models, video surveillance, drone and or helicopter operations, implementations of community police networks & neighborhood watches, social-media watch units, Automatic Tag and License Plate Readers, gunshot detection systems. These could be measured and included in a dataset that corresponds to degree of utilization, geographic coverage, population coverage, funding, and so forth. All of our authority variables would be hypothesized to have a negative relationship with our response variable.

Our final category that could be aided by additional variables is that of geography. For example, omitted variables such as the number of locations (or density thereof) of casinos, bars, pawn-shops, and other licensed areas that are typically associated with criminal activity could contribute positively to *crime*. Further geographic variables such as the number of abandoned/condemned houses or other NIMBY (not in my backyard) features could add to a location's crime.
Two general concern of the researchers emerge from the use of counties as the primary geographic unit of differentiation for each observation. On the one hand, each county could potentially represent too large of a region, particularly if the county has many different cities and municipalities that have different levels of crime. Aggregated together, these might present a distorted picture. This could be helped by the *urban* variable, as it generally indicates whether a significant portion of the county might be affiliated with a city but there could still be significant portions of a county that differ dramatically from other portions of the county.

On the other hand, we are presented with the concern that individuals, particularly when they live on the boarder of a county or have a commute, might confound our observations as individuals who are counted as residents in one county might commit all of their crimes in another county. Consequently, larger areas that are more targeted towards the different specific metropolitan statistical areas might better be used.

As a function of our lack of control over the collection of our observations, we consequently are very limited in our ability to discuss causality to the degree that we are not comfortable assuming any causality due to the significant shortcomings of the data. For example, the data and our model might be facially incorrect, as our response variable, *crime*, might be inadequately measured. Indeed there is probably a high likelihood that there is a combination of misreporting and underreporting of the *crime* variable due to the nature of that which is being reported - criminal activity and incidents. It is common knowledge that in some neighborhoods and communities, sentiment towards the government and law enforcement is decidedly negative, and consequently individuals might have a disincentive to report criminal activity, suspected or confirmed. To definitively prove causality among our thesis, variables, and models, it would be necessary to conduct Randomized Control Trials, ideally in a Solomon 4-group protocol and observe all aspects of the experiment design, execution, and data gathering to ensure that no bias or data is negatively impacted. Because the feasibility of this is rather low, we would propose the development of a series of field experiments, as controlled as possible to ascertain the drivers and levers that impact *crime*.

Conclusion

As our model and analyses are associative as opposed causal, without further study, we are unable to definitively articulate a causal plan to reduce crime. Our initial hypothesis of creating a thesis around a "carrot and stick" proved unfortunately futile in finding a silver bullet, though with more robust data, this thesis may hold true. However, our modeling suggests a number of intermediary conclusions. We are able to draw a conclusion from our second model that the percentage of young adult males and density of a county positive contribute crime with a significant effect size at a statistically significant level, while the percentage of population that are minorities and the average wage of public sector employees also are statistically significant, but do not cause much of a change in our response variable, *crime*. We are also able to conclude that police presence per capita, although not statistically significant, potentially does have a serious impact on crime in a practical significance sense.

Consequently, for policy and programmatic considerations, additional focus of resources and attention should be devoted to areas where there is a high number of young adult males and high population density in general. We advocate that these two areas should be investigated further in addition to a more comprehensive experiment about the drivers of crime; we can only prognosticate as to why these factors play a role in crime and consequently what further steps should be taken within communities with high percentages of young adult males and of high density. As a result of the limitations on our data set, associated metadata and protocol used to collect the data, we are constrained in our ability to deliver a robust model and analysis that we initially set out and wished to complete.

Works Cited

Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2. http://CRAN.R-project.org/package=stargazer