Lab 4: INFORMATIVE TITLE

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 dataset 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 apporach, 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

```
setwd("C:\\Users\\Krissy\\Dropbox\\W203 - Statistics\\Lab 4")
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 codebook.

```
# summary(full_data)
# CAN WE USE COLNAMES HERE INSTEAD? SUMMARY JUST OUTPUTS A LOT. BETTER TO USE THAT OUTPUT LOWER DOWN, W
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"
```

From our high-level summary statistics and descriptions from the codebook, we observe that some variables are categorial-ordinal (county, year), coded (west, central, urban), proportions or probabilities (probarr,

[1] https://en.wikipedia.org/wiki/Carrot_and_stick

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 reliabile 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 codebook 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 identifer". 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

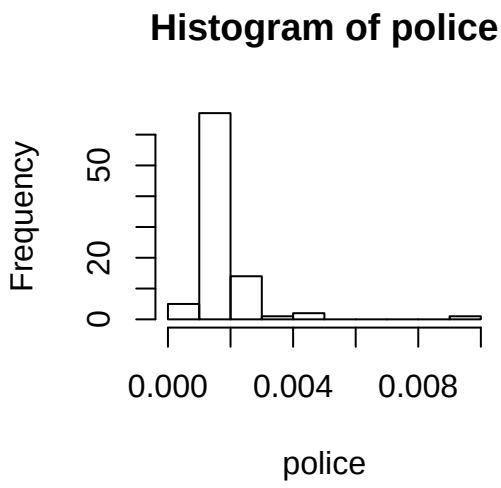
Authority is _____.

We have identified is the police variable, described as "police per capita" from the code book.

```
police <- data$police
summary(police)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0007459 0.0012378 0.0014897 0.0017080 0.0018856 0.0090543
```

We observe that our police per capita rate depicts that there is on average, across all observed counties, 1.7 police personnel per 1000 individuals and there is a significant positive skew to the distributrion.



Demographics

We are interpereting demographics as _____.

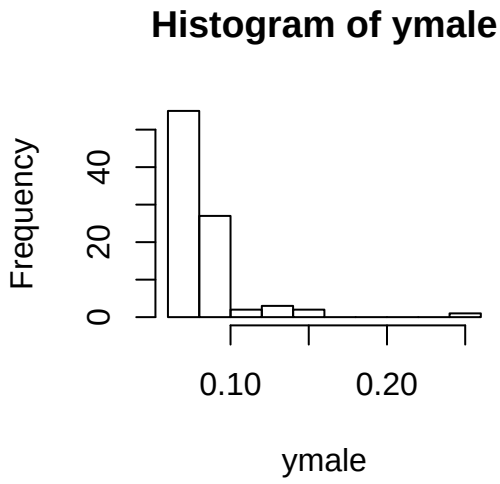
There are two variables in the demographics category, ymale and pctmin, or proportion of county males between ages 15 and 24, and proportion that is minority or nonwhite, respectively.

Our primary variable of analysis for Demographics will be ymale, as other studies have demonstrated that young males have, on average, a higher propensity to commit crimes.

```
ymale <- data$ymale
summary(ymale)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06216 0.07437 0.07770 0.08403 0.08352 0.24871
```

The ymale data depict an average of 8.4% young males per county. We further observe a strong positive skew with one value reaching nearly 1 in 4 residents is classifed as a "young male."

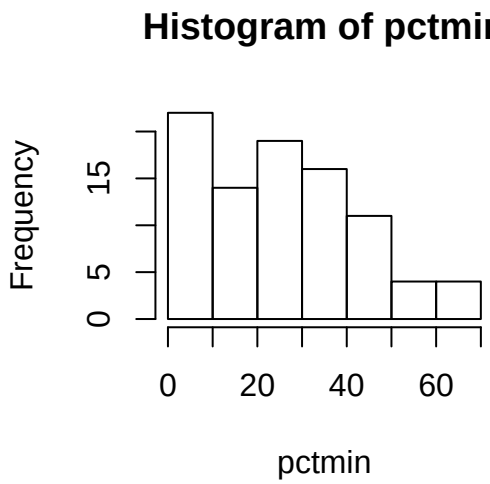


An additional variable that ilustrates demographics is that of pctmin, or the percentage of minorities in each county.

```
pctmin <- data$pctmin
summary(pctmin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.284  10.024  24.852  25.713  38.183  64.348
```

We observe that there is a wide distribution of this variable, ranging from 1.2% to 64.3% Of note is that this data format is not congruent with our previous ymale format, wheras this depicts integer-percents, as the other depicts decimal-percents.



From the histogram plot, we see a positively-skewed distribution.

Geography

Geography in our model represents _____.

There are a number of variables we have identified that relate to geography, including density (people per sq. mile), west and central, binary variables illustrating whether the observed county is in the west or central portion of the state, and urban, another binary variable representing if the county is in the standard metropolitan statistical area.

Our two primary geography variables are density, and urban.

Our secondary variables relate to the location of the county and is a combination of west and central variables.

Economics

Finally, economics is in our model as _____.

Included in this group of variables include all the weekly wage variables (wagecon (Construction), wagetuc (Transportation, Utilities, Communciations), wagetrd, (wholesale, retail trade), wagefir (finance, insurance and real estate), wagemfg (manufacturing), wagefed (Federal Employees), wagesta (state employees), and wageloc (local government employees)). Finally we have also categorized tax, or the tax revenue per capita as an economics variable.

Our primary economics variable is the _____.

Our secondary economics variable is tax, or tax revenue per capita. We have demoted this variable to our secondary analysis because of the limitations in understanding the parameters of this variable. We are concerned that without further information, the tax variable might comingle differnet types of tax data, representative of not only personal taxes, but also taxes of businessess. Consequently we observe this variable with a discerning and suspicious eye, and take it's contours and parameters with a grain of salt.

WARNING: Because we don't know the population of the counties or denominator of the average wages, the "per capita/averages" could be signficantly skewed or not reliably comparable across all counties. We nevertheless proceed with caution.

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.
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.

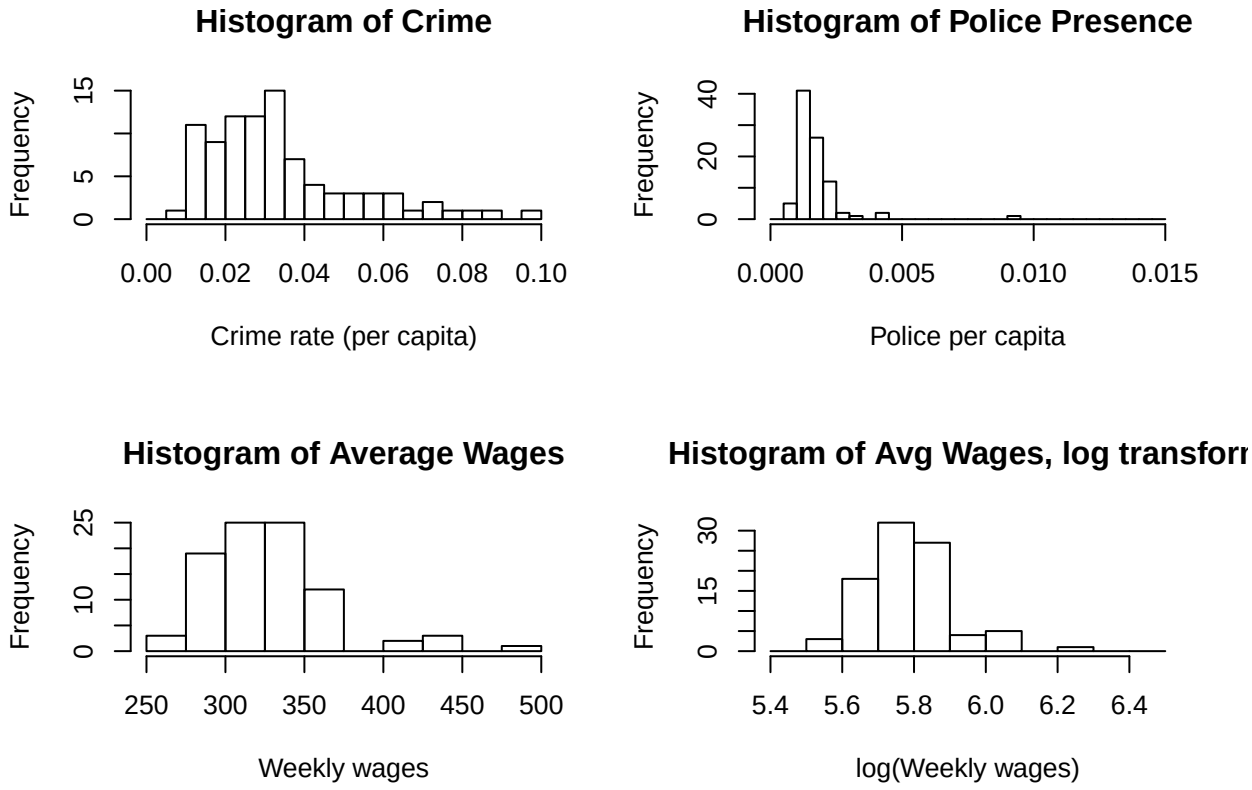
Model 1:

Specification

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

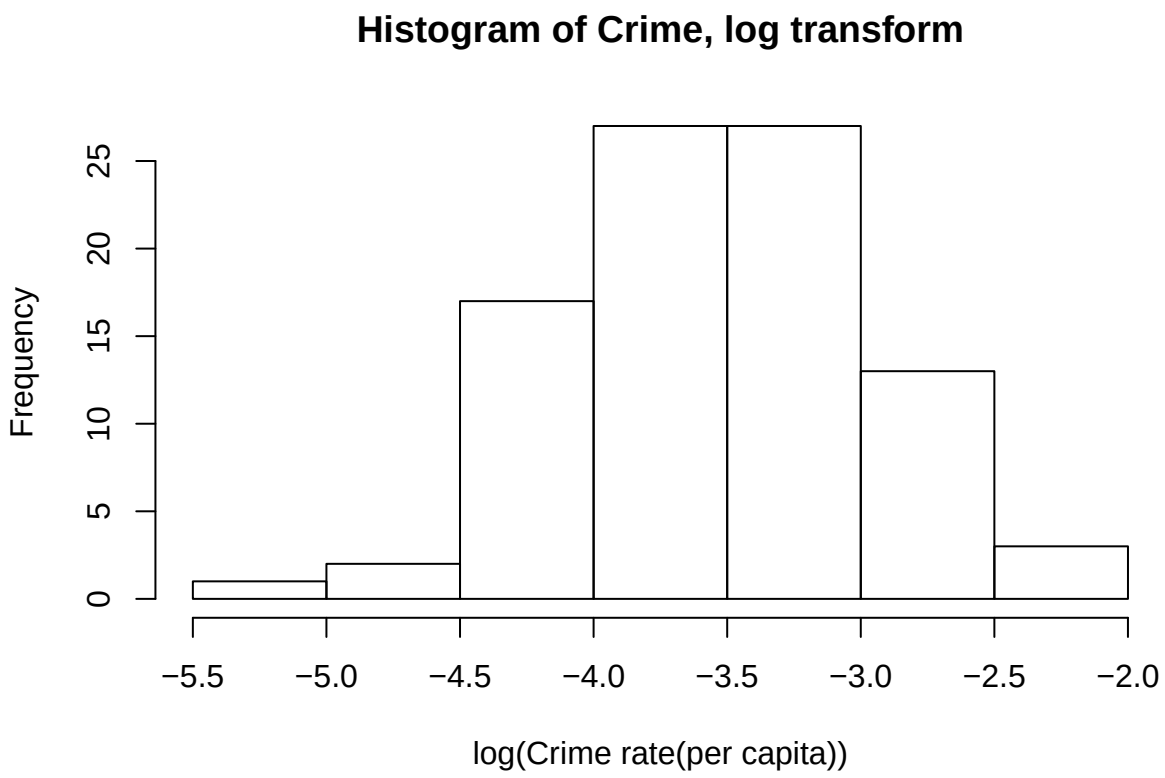
Histograms of each of these variables allow us to identify skew and make any required transformations. NOTE: It is quite common to transform salary/wage variables using a logarithm. In anticipation of that, a histogram of *log(wage_avg)* is provided proactively here.

```
# Create the variable for average wage
wagevars = c("wagefed", "wagesta", "wageloc", "wagecon", "wagetuc", "wagetrd", "wagefir", "wageser", "wa
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

hist(data$police, main = "Histogram of Police Presence", xlab = "Police per capita", breaks = seq(0,0.0

hist(data$wage_avg, main = "Histogram of Average Wages", xlab = "Weekly wages", breaks = seq(250, 500, 1

# Create a log(wage) variable
data$log_wage_avg <- log(data$wage_avg)
hist(data$log_wage_avg, main = "Histogram of Avg Wages, log transform", xlab = "log(Weekly wages)", brea
```



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.

```r
# 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 policeman 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:

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

```r
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 our data fits our model assumptions. As our model uses a “carrot” and “stick” approach we investigate the former and later variables. We identify that police presense (“stick”) is appropriate for our model, but our “carrot” is potentially of concern as we are taking the “average of averages” in this context.

We are presented with no other data and consequently will proceed with the understanding that this may limit the full interpretability of our model.

We are able to pass the first assumption, MLR.1 as we have developed our model into a linear form as described supra.

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 with the assumption that our model meets MLR.2.

The third assumption, MLR.3, is 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:

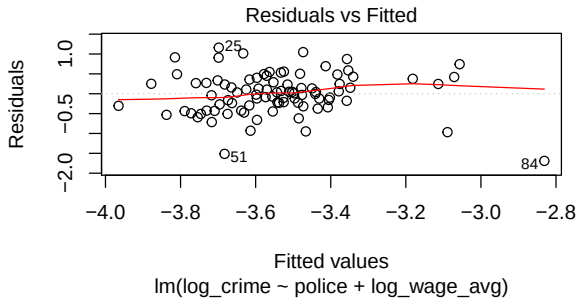
```
(VIF1 <- 1/tolerance1)
```

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

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

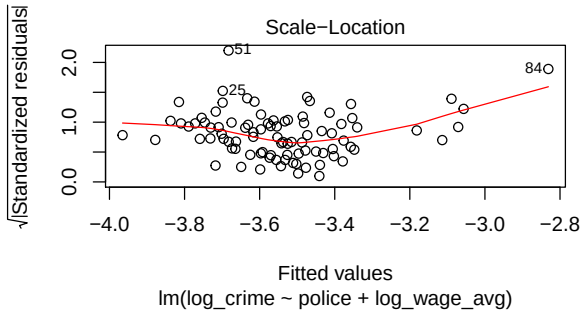
The next assumption we examin is that of the Zero Conditional Mean, MLR.4. We first identify if there are any omitted variables - as we have a limited dataset with respect to different variables, we are caught in a bit of a bind. Theoretically there are many more "carrot" and "stick" variables that we 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 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, we can observe that the band has a roughly uniform thickness, with the exception of points 51 and 84. We will proceed cautiously with further analysis of homoskedasticity.



Our scale-location plot also depicts a relatively horizontal band of points, with the exception of point 84 on the right.

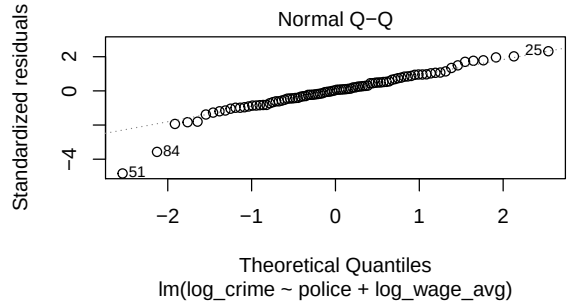
```
bptest(Model1)
```

```
##
##  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-robuts methods of analysis going forward. (See the "Summary of Models" section.)

We proced with the analysis of MLR.6, The normality of errors.

We first investigate the residual plot, which is roughly normally distributed:



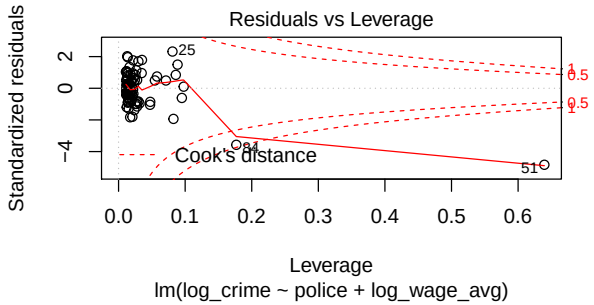
Our QQ test shows that the majority of points falls on the diagonal line with some departure at the lower

extreme - again in points 51 and 84.

```
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),]
```

```
##        crime avgsen     police   density      tax west central urban
## 51 0.0055332  20.70 0.00905433 0.3858093 28.19310    0       1     0
## 84 0.0108703   5.38 0.00122210 0.3887588 40.82454    1       0     0
##      pctmin  wagecon  wagetuc  wagetrd  wagefir   wageser wagemfg wagefed
## 51  1.28365 204.2206 503.2351 217.4908 342.4658  245.2061  448.42  442.20
## 84 64.34820 226.8245 331.5650 167.3726 264.4231 2177.0681  247.72  381.33
##    wagesta wageloc      ymale wage_avg log_wage_avg log_crime
## 51  340.39  386.12 0.07253495 347.7498     5.851483 -5.196989
## 84  367.25  300.13 0.07008217 495.9648     6.206505 -4.521721
```

```
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 the third lowest crime rate (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, we can rely on asymptotic properties of OLS including the central limit theorem.

Model 2:

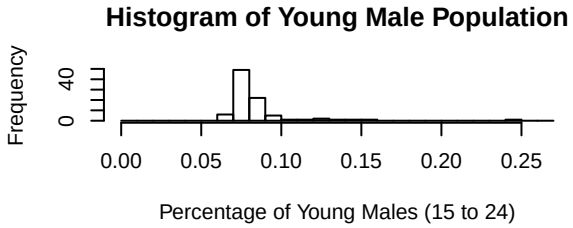
Specification

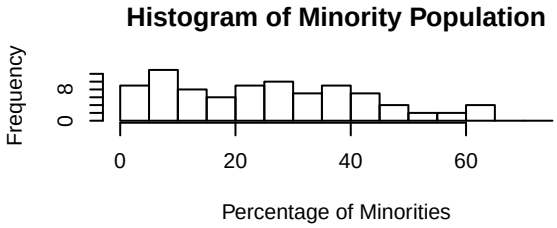
This model adds a bit of complexity to the first, in order to better capture the full range of variables that affect crime. We incorporate interaction terms and covariates where they might affect the model. In particular, we include at least one variable from each category discussed above, in order to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

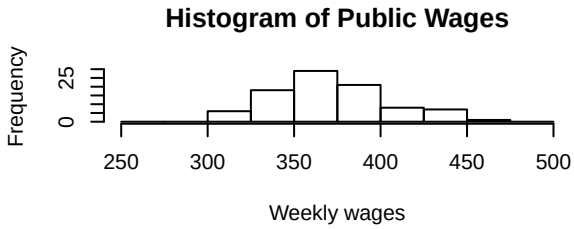
- Demographics: operationalized by the number of young males in the county, *ymale*, and the percentage of minority citizens, *pctmin*
- Geography: operationalized by population density, *density*
- Economics: in place of the incentive variable *wage_avg* captured in Model 1, the average wages of government 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*).

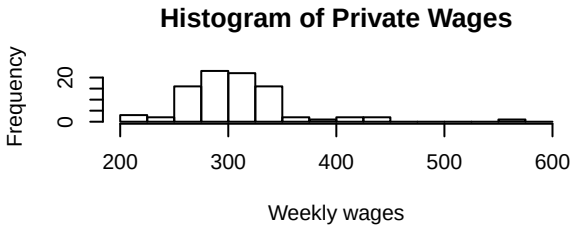
Histograms of each of these variables allow us to identify skew and make any required transformations.

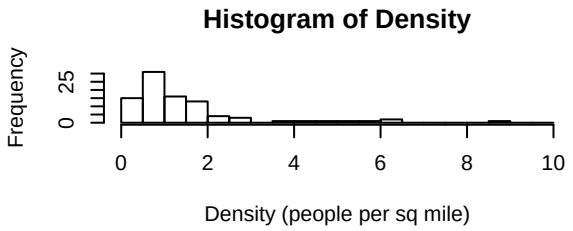
```
# Create the variables for public & private wages
pubwages = c("wagefed", "wagesta", "wageloc")
privwages = c("wagecon", "wagetuc", "wagetrd", "wagefir", "wageser", "wagemfg")
allwages = c(pubwages, privwages)
data$wage_public = apply(data[,pubwages], 1, mean)
data$wage_private = apply(data[,privwages],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

hist(data$pctmin, main = "Histogram of Minority Population", xlab = "Percentage of Minorities", breaks =

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

hist(data$wage_private, main = "Histogram of Private Wages", xlab = "Weekly wages", breaks = seq(200, 6

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











The two independent variables *ymale* and *density* 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:

$$log(Crime) = \beta_0 + \beta_1 * Police + \beta_2 * Ymale + \beta_3 * PctMinority + \beta_4 * Density + \beta_5 * PublicWages + \beta_6 * PrivateWages$$

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

Assumptions

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

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

```
(tolerance2 <- 1-summary(Model1)$r.squared)
```

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

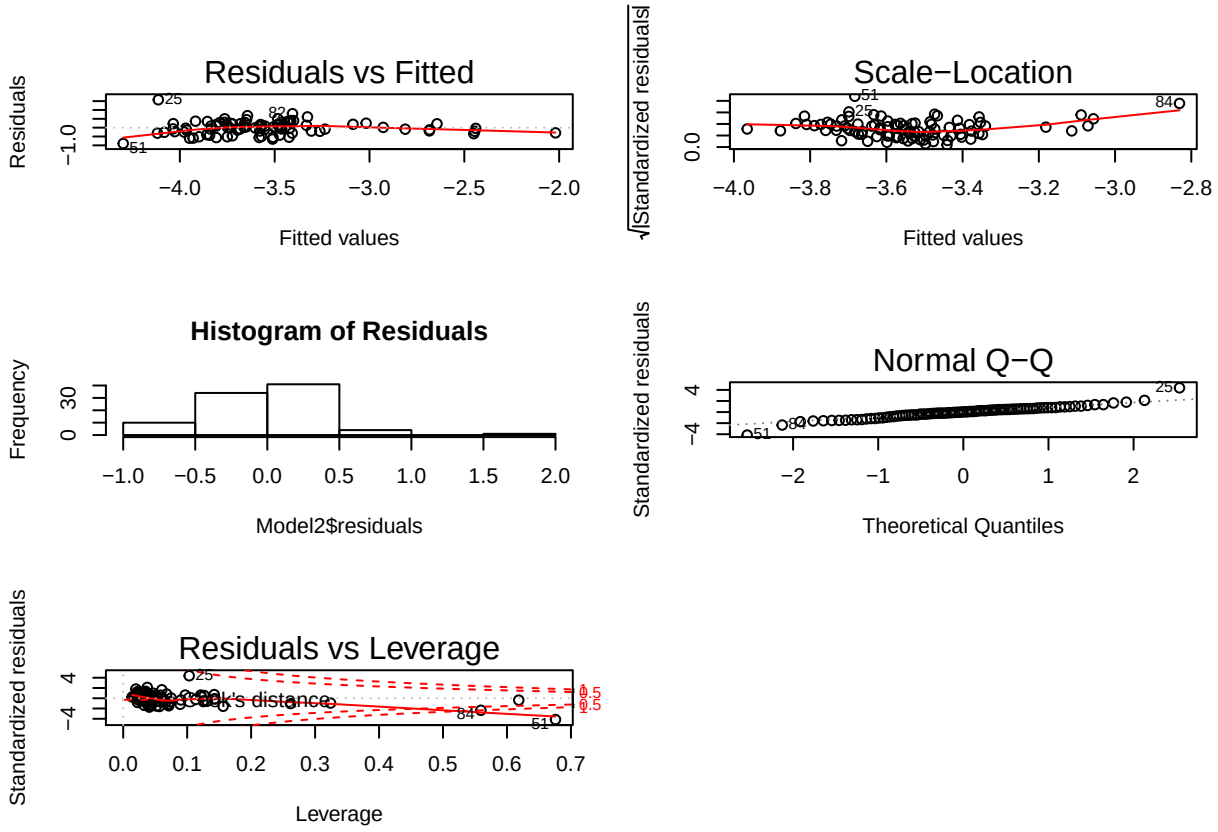
```
(VIF2 <- 1/tolerance2)
```

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

```
bptest(Model2)
```

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

```
#Diagnostic Plots
par(mfrow=c(3,2))

# Residuals vs Fitted Values
plot(Model2, 1)

# Scale-Location
plot(Model1, 3)

# Histogram of residuals
hist(Model2$residuals, main = "Histogram of Residuals")
# Normal Q-Q
plot(Model2, 2)

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



ITS POINTS 51 AND 84 AGAIN CAUSING ISSUES. WE ALREADY KNEW THAT. IN TERMS OF

CHANGES FROM MODEL 1, MODEL 2 HAS AN INPROVED R^2 VALUE AND GENERALLY CLEANER DIAGNOSTICS.

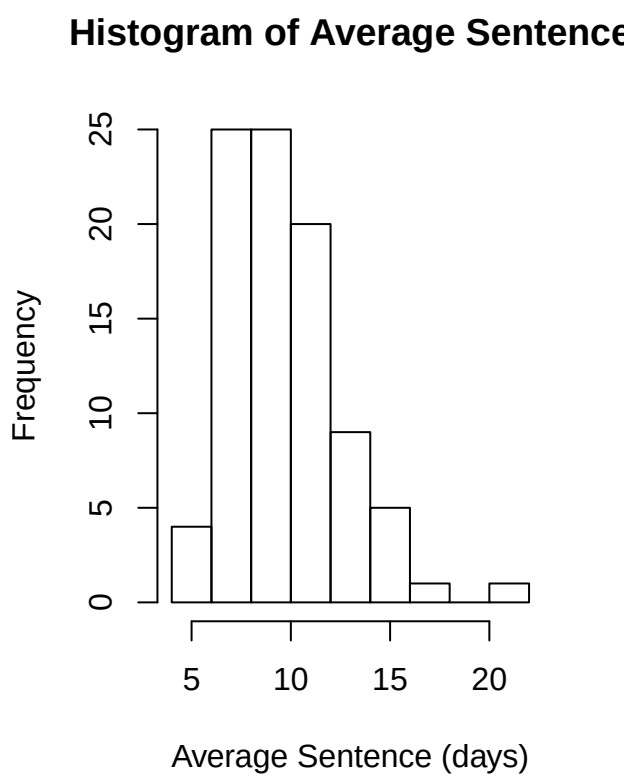
Model 3:

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.

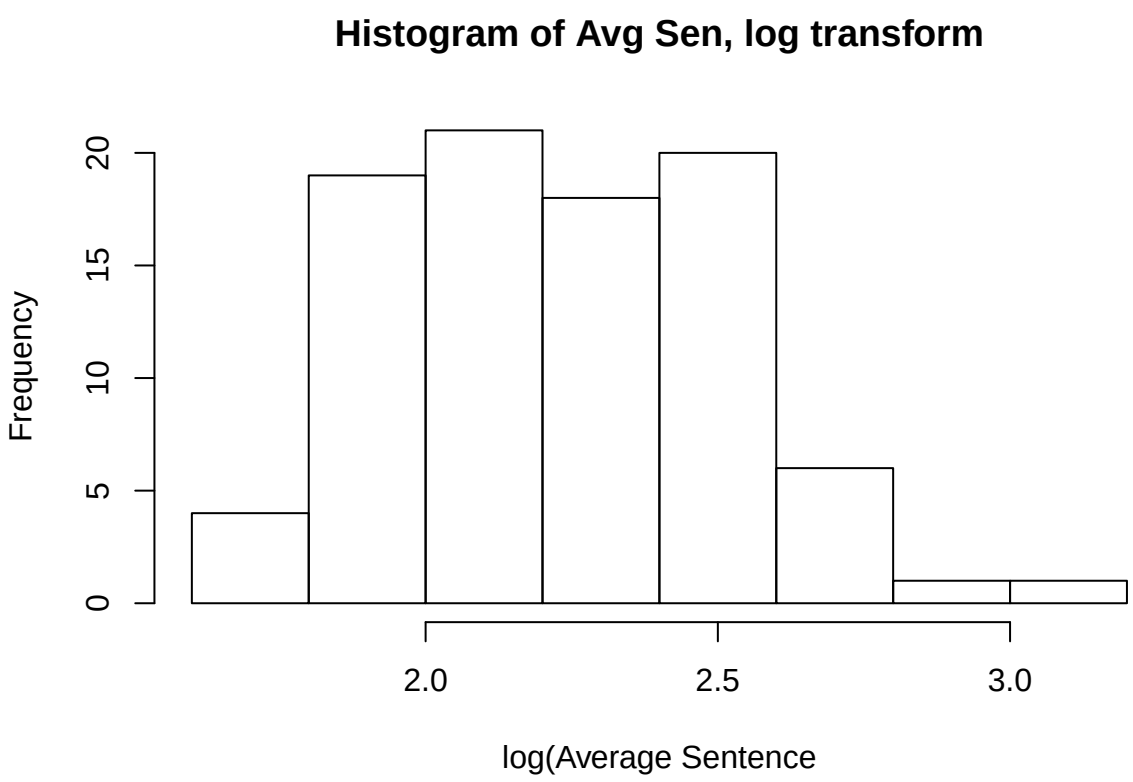
A histogram of the *avgsen* variable allows us to identify skew and make any required transformations.

```
par(mfrow=c(1,2))

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



We should consider 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:

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

```
Model3 <- lm(log_crime ~ police + ymale + pctmin + density + wage_public + wage_private + urban + urban

summary(Model3)
```

```
##
## Call:
## lm(formula = log_crime ~ police + ymale + pctmin + density +
##     wage_public + wage_private + urban + urban * ymale + log(avgsen),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89756 -0.22186  0.00853  0.20013  1.59338
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.133e+00  6.520e-01  -7.873 1.43e-11 ***
## police       -3.409e+01  4.606e+01  -0.740 0.461325
## ymale         4.243e+00  1.844e+00   2.301 0.024007 *
## pctmin        8.731e-03  2.502e-03   3.490 0.000789 ***
## density       2.641e-01  5.398e-02   4.893 5.05e-06 ***
## wage_public   3.430e-03  1.690e-03   2.029 0.045803 *
## wage_private -1.557e-03  9.662e-04  -1.611 0.111029
```

```
## urban        -8.156e-01  7.082e-01  -1.152 0.252921
## log(avgsen)  -3.479e-02  1.646e-01  -0.211 0.833128
## ymale:urban   5.528e+00  6.861e+00   0.806 0.422780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3824 on 80 degrees of freedom
## Multiple R-squared:  0.5636, Adjusted R-squared:  0.5145
## F-statistic: 11.48 on 9 and 80 DF,  p-value: 2.386e-11
```

Assumptions

Summary of Models

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

NICE! MODEL2 HAS LOWEST AIC — MOST PARSIMONIOUS BEST FIT =D

Significance

Statistical Significance

Practical Significance

INTERESTING THAT DENSITY SEEMS MOST SIGNIFICANT
POLICY-WISE, PERHAPS THE "CARROT" AND THE "STICK" *BOTH* DON'T MATTER. DON'T NEED TO MANIPULATE THINGS THAT SO CLEARLY TARGET CRIME. INSTEAD, FOCUS ON ZONING AND OTHER GEOGRAPHIC INFLUENCES. BONUS EFFECT OF MAKING THE CITIZENS FEEL LESS CONTROLLED/WATCHED/BIG-BROTHER-ED.
ALSO, PUBLIC WAGES ARE MORE OF AN ISSUE THAN PRIVATE WAGES. GOVERNMENT SHOULD NOT TAKE HIGHER PAY?

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

Causality & Omitted Variables

Conclusion

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