Lab 3: Reducing Crime (DRAFT: Stage 1)

C. Akkineni, A. Thorp, K. Hanna

November 27, 2018

Contents

Introduction (Stage 1: Draft Report)

The team has been hired to provide research for a political campaign and help the campaign understand the determinants of crime and to help with policy suggestions that are applicable to local government.

```
library(knitr)
library(kableExtra)

codebook <- read.csv('codebook.csv')
crime <- read.csv('crime_v2.csv')

# Convert columns to factors and logical.
crime$county <- as.factor(crime$county)
crime$year <- as.factor(crime$year)
crime$west <- as.logical(crime$west)
crime$central <- as.logical(crime$central)
crime$urban <- as.logical(crime$urban)

# Create a log of the dependent variable
crime$logcrmrte <- log(crime$crmrte)
```

Exploratory Data Analysis

Data Summary

We were provided with a dataset of crime statistics for a selection of counties in North Carolina. After performing data clean up (outlined below) the data set contained 90 county observations each having 25 variables (onlined in the codebook found in Appendix A).

Data Clean Up

Null Rows

The dataset contained a an apostrophe 6 rows after the data which caused the csv reader to create 6 invalid rows. We feel it is safe to remove these rows as they contain no data.

```
# Delete the 6 empty observations at the end, including the row with the apostrophe.
# We can use complete.cases to do this as these 6 observations are the only incomplete observations.
crime = crime[complete.cases(crime), ]

# Fix prbconv which is a factor rather than numeric due to the apostrophe
# Convert from factor to numeric
crime$prbconv = as.numeric(as.character(crime$prbconv))
```

We found two identical observations for county 193. There is no logical reason to have two identical observations in this cross-sectional data, so we feel strongly that removing one of these two observations can only benefit our analysis.

```
# county 193 is duplidated, remove one
crime = crime[!duplicated(crime), ]
```

Concerns about data

There are three probability columns in the given dataset. Check if any of the columns has invalid values - i.e., any of the columns have less than zero or greater than 1 values.

```
#any(crime$prbarr<0 | crime$prbarr>1)
#any(crime$prbconv<0 | crime$prbconv>1)
#any(crime$prbpris<0 | crime$prbpris>1)

#summary(crime$prbarr)
#summary(crime$prbconv)


summary(crime$prbarr)[c(1,6)]
```

```
##    Min.    Max.
## 0.09277 1.09091
```

```
summary(crime$prbconv)[c(1,6)]
```

```
##      Min.      Max.
## 0.0683761 2.1212101
```

```
summary(crime$prbpris)[c(1,6)]
```

```
## Min. Max.
## 0.15 0.60
```

```
nrow(crime[(crime$prbarr<0 | crime$prbarr>1), c('county', 'prbarr')])
```

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

```
nrow(crime[(crime$prbconv<0 | crime$prbconv>1), c('county', 'prbconv')])
```

```
## [1] 10
```

prbarr (Probability of Arrest)

We found that county 115 contained a value of 1.09 in prbarr (probability of arrest) which is not possible. It is the only observation with an invalid value, we will treat this variable as valid, however do so with caution.

prbconv (Probability of Conviction)

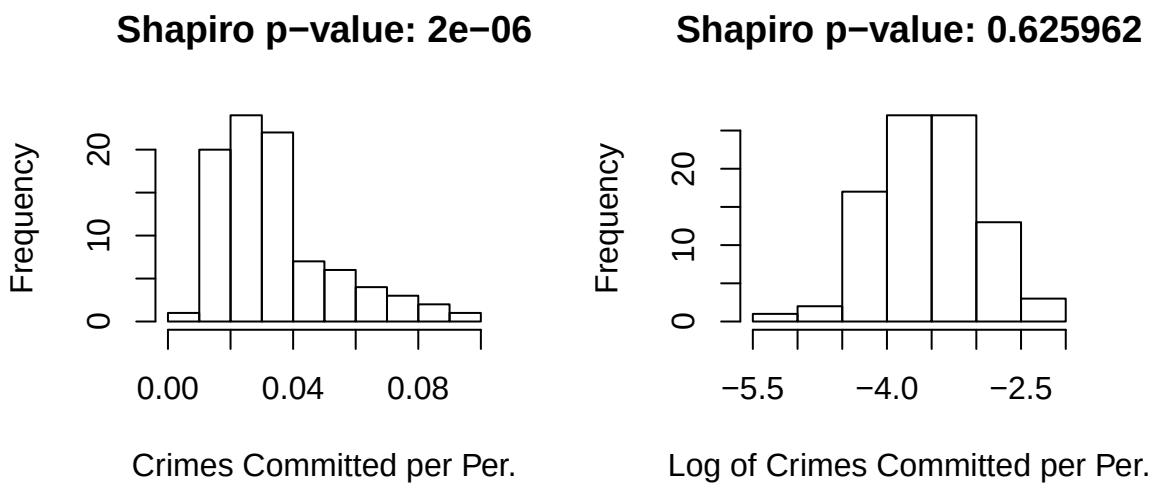
We found 10 observations with values greater than 1, which, again, is not a possible value for probability. We have little confidence in the veracity of this variable, and will not be performing further analysis.

Univariate Analysis

```
quick_uni_analysis = function(variable, description) {
  hist(variable, xlab = tools::toTitleCase(description),
     main = paste('Shapiro p-value:',
                  round(as.numeric(shapiro.test(variable)[2]), 6)
                  )
     )
  hist(log(crime$crmrte),
       xlab = tools::toTitleCase(paste('Log of', description)),
```

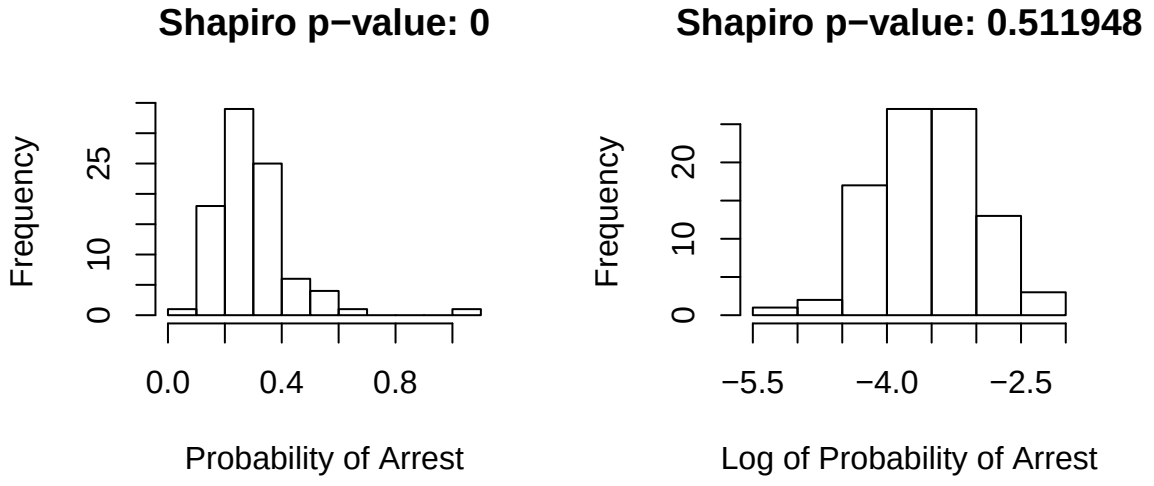
```
        main = paste('Shapiro p-value:',
                     round(as.numeric(shapiro.test(log(variable))[2]), 6)
                     )
        )
}
```

```
quick_uni_analysis(crime$crmrte, 'crimes committed per per.')
```



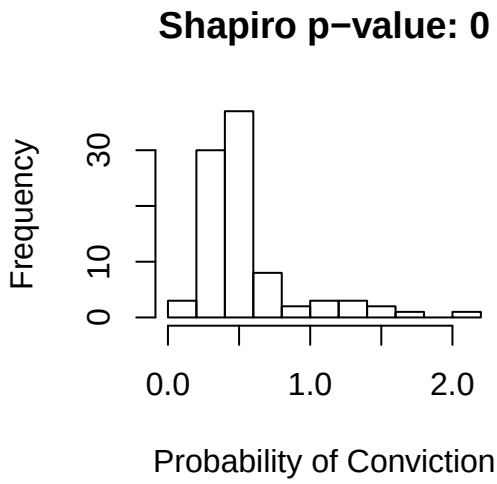
Log is preferable - both for interpretation and for better adhering to modeling assumptions

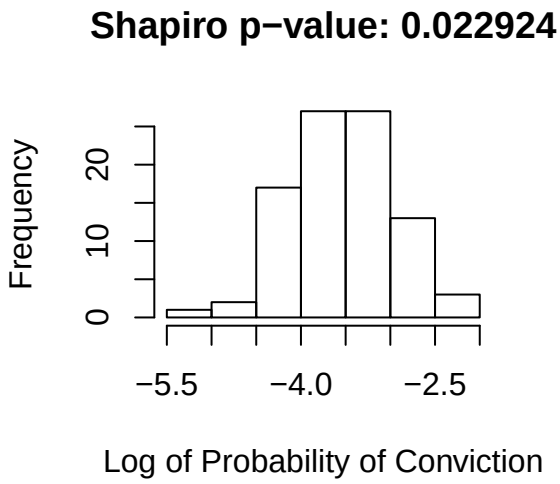
```
quick_uni_analysis(crime$prbarr, 'Probability of Arrest')
```



Log is preferable - both for interpretation and for better adhering to modeling assumptions

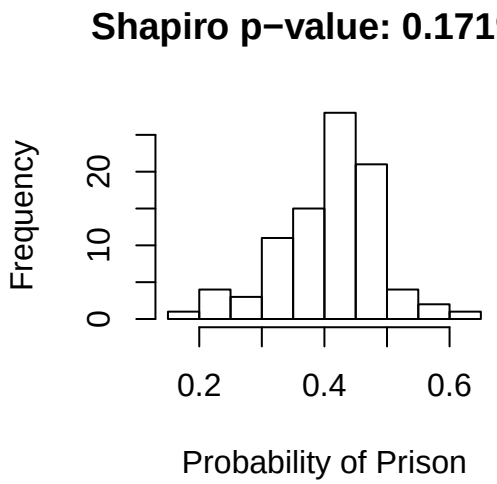
```
quick_uni_analysis(crime$prbconv, 'Probability of Conviction')
```

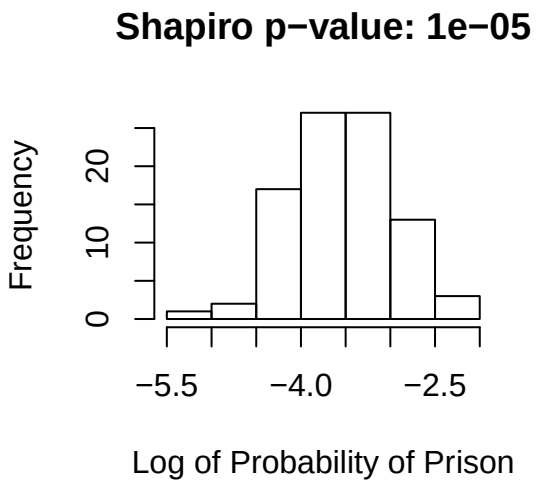




Log is preferable - both for interpretation and for better adhering to modeling assumptions. However, even the logged version fails a Shapiro-Wilk normality test. Something to keep in mind.

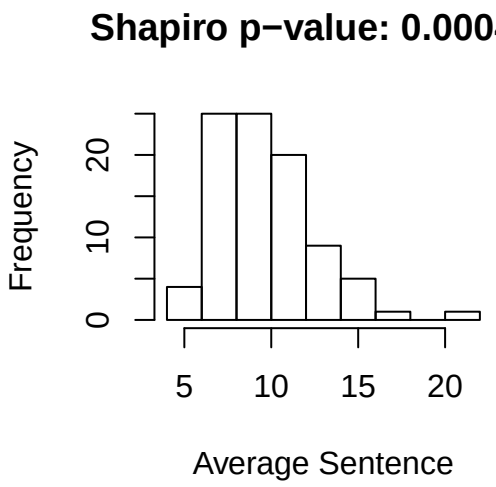
```
quick_uni_analysis(crime$prbpris, 'Probability of Prison')
```

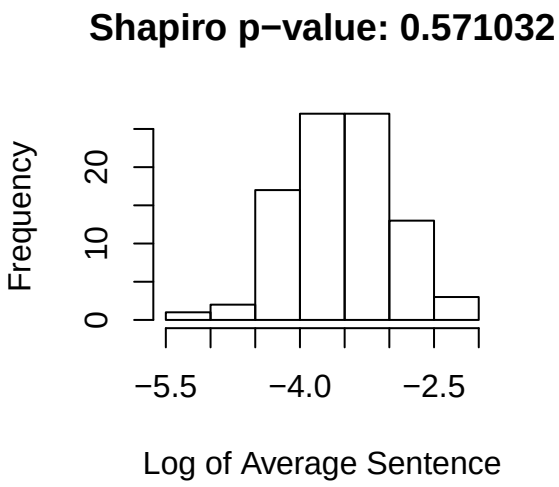




From an interpretation standpoint, the logged version is preferable, although from an modeling assumption standpoint, the unlogged version is preferable.

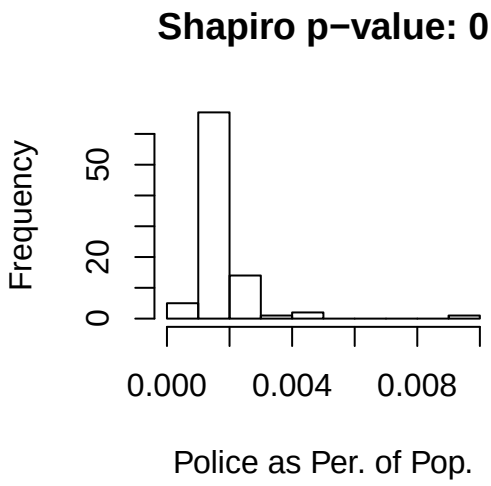
```
quick_uni_analysis(crime$avgsen, 'Average Sentence')
```

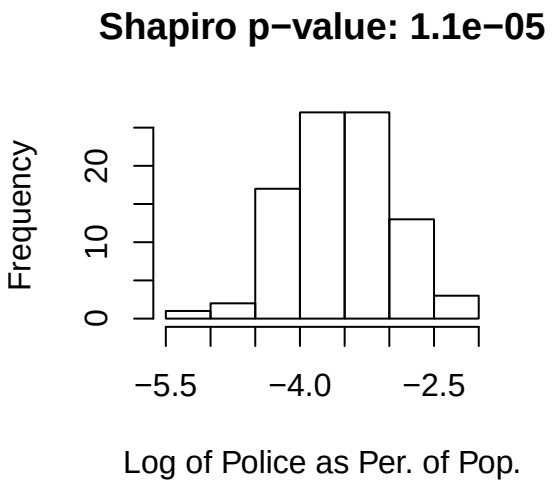




The logged version is preferable from both an interpretation and modeling assumption standpoint.

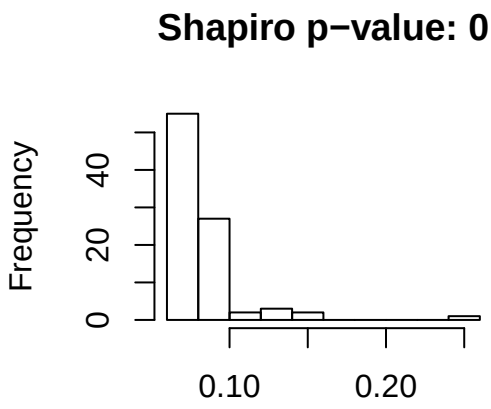
```
quick_uni_analysis(crime$polpc, 'Police as Per. of Pop.')
```



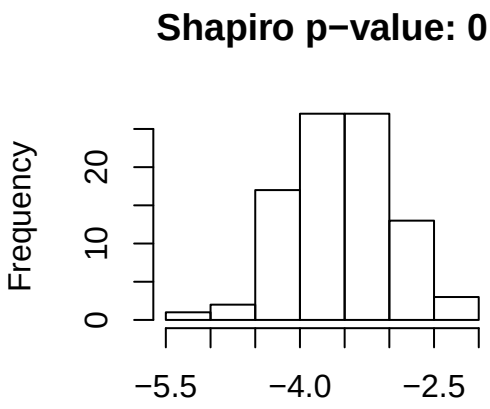


Both logged and un-logged versions of police as a percentage of the population are non-normal. Neither is inherently preferable from a modeling assumptions standpoint.

```
quick_uni_analysis(crime$pctymle, 'Per. of Pop. That Are Young Males')
```



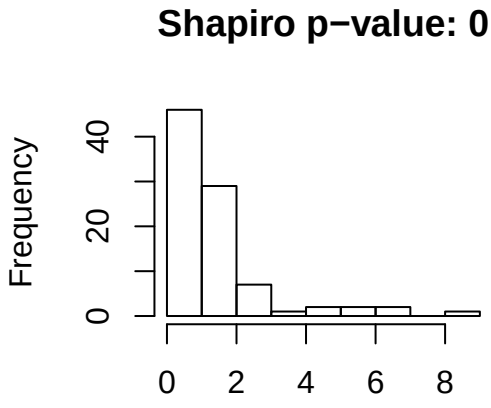
Per. of Pop. that are Young Males



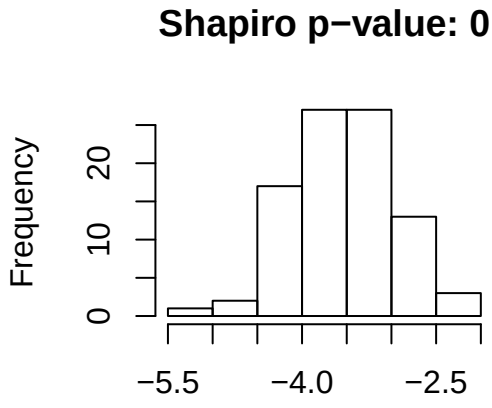
Log of Per. of Pop. that are Young Mal

Both logged and un-logged versions of the percent of population that is young and male are non-normal. Neither is inherently preferable from a modeling assumptions standpoint.

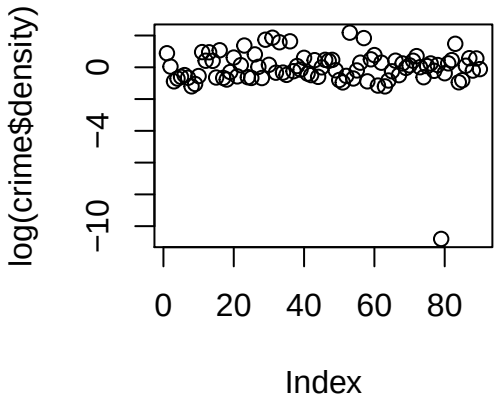
```
quick_uni_analysis(crime$density, 'people per sq. mile')
plot(log(crime$density))
boxplot(log(crime$density))
plot(crime$logcrmrte)
boxplot(crime$logcrmrte)
```

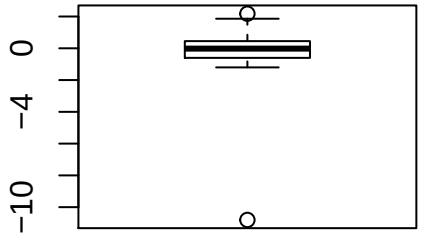


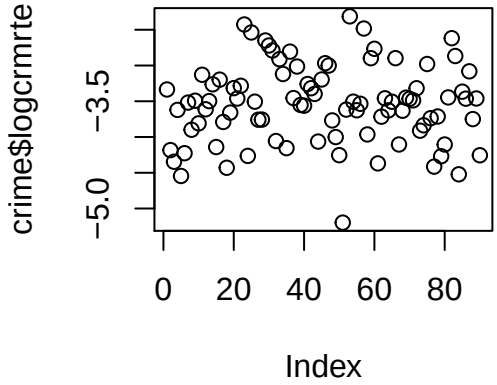
People per Sq. Mile

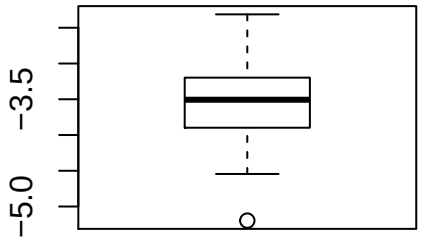


Log of People per Sq. Mile


log(crime$density)
−10
−4
0
0
20
40
60
80
Index


−10
−4
0


crime$logcrmrte
−5.0
−3.5
0
20
40
60
80
Index


−5.0
−3.5

Model Analysis

General Crime Prediction Model

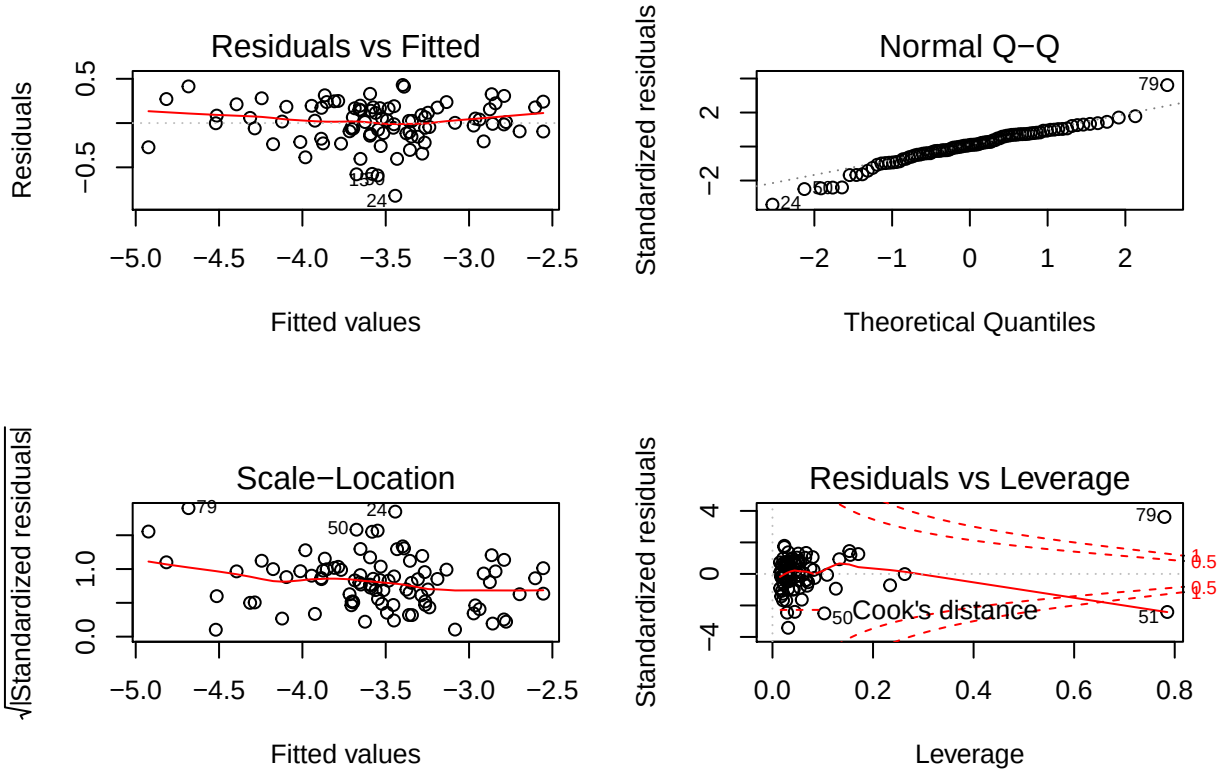
```
general_model = lm(logcrmrte ~ prbarr + prbconv + polpc + log(pctmin80) + log(density), data = crime)
general_aic = AIC(general_model)
general_rsquared = summary(general_model)[8]
general_adjrsquared = summary(general_model)[9]
print(general_aic)
```

```
## [1] 9.393001
```

```
print(general_adjrsquared)
```

```
## $adj.r.squared
## [1] 0.8020671
```

```
par(mfrow = c(2,2))
plot(general_model)
```




```
#plot(general_model, which = 1)
#plot(general_model, which = 5)
```

Preliminary Infomation (not intended to be left in)

From the assignment:

- 1. What do you want to measure? Make sure you identify variables that will be relevant to the concerns of the political campaign.
- 2. What transformations should you apply to each variable? This is very important because transformations can reveal linearities in the data, make our results relevant, or help us meet model assumptions.
- 3. Are your choices supported by EDA? You will likely start with some general EDA to detect anomalies (missing values, top-coded variables, etc.). From then on, your EDA should be interspersed with your model building. Use visual tools to guide your decisions.
- 4. What covariates help you identify a causal effect? What covariates are problematic, either due to multicollinearity, or because they will absorb some of a causal effect you want to measure?

Variables:

1. Target
- crmrte
2. Label
- county
3. Geographic: (my own word, just things that can segregate the data). Counties can belong to 0 or more from west, central and urban.
- density (likely related to others, especially urban)
- west
- central
- urban
4. Certainty of Punishment:

- prbarr
- prbconv

5. Severity of Punishment:

- prbpris
- avgsen

Other + polpc (likely related to certainty of punishment)

We haven't really addressed what kinds of crimes are being talked about. Should we assume that it is violent crimes and drug crimes? (i.e. not white collar or other non-violent) If we also assume that some crimes are committed by rational actors weighing gains versus losses, we would expect crimes like theft and drug dealing to be negatively correlated with wealth (i.e. higher opportunity cost for wealthy people to engage in those sorts of crimes - and by extension linked crimes like manslaughter (I'm thinking an attempted theft or drug deal gone wrong).

The idea here is that higher real wages reduce crime propensity. To operationalize that, ideally we would have some idea of cost of living and nominal wages (particularly in sectors that are potential alternatives to those who typically engage in violent crime and drug dealing - those without high educational barriers to entry - construction, driving, retail).

In addition to pure wealth, there are other ideas like social capital and a sense of belonging/rootedness that we can imagine would be associated with lower propensity to commit crimes. We could assess this through variables like homeownership, rate of volunteering, rates of religious attendance, or the proportion of mothers that are single and have sole custody.

Another thought is that educated people are also less likely to committ these kinds of crimes. Another relevant variable may be average years of education, proportion of population with at least a high school diploma, at least a bachelor's, or one or more graduate degrees. Related to this category as well as the previous notion of social belonging, the rate of absenteeism among high school students could also reflect these two notions.

Another idea is that proximity to instruments of crime also increases propensity to commit crimes. The rate of gun ownership, or - even better - the rate of illicit gun ownership would be useful here.

Another idea is that crimes are more likely as inequality is high between neighboring locations - creating a threshold upon which economic and extension crimes become focused. This is hard to assess without better location data

And another idea is that organized crime tends to be more persistent than isolated instances of crime. So, to the extent possible, it would be useful to have a measure of gang/mob/etc membership as a percent of the population.

```
mod1 <- lm(crmrte ~ prbarr + prbconv + prbpris + avgsen, data=crime)
summary(mod1)
```

```
##
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + prbpris + avgsen, data = crime)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.031434 -0.009597 -0.002289  0.007937  0.052989
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0446704  0.0109773   4.069 0.000105 ***
## prbarr      -0.0625420  0.0121899  -5.131 1.80e-06 ***
## prbconv     -0.0234739  0.0047115  -4.982 3.27e-06 ***
## prbpris      0.0212198  0.0204908   1.036 0.303336
## avgsen       0.0011892  0.0006007   1.980 0.050992 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01549 on 85 degrees of freedom
## Multiple R-squared:  0.358,  Adjusted R-squared:  0.3278
## F-statistic: 11.85 on 4 and 85 DF,  p-value: 1.072e-07
```

```
mod1_log <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgsen), data=crime)
summary(mod1_log)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##     log(avgsen), data = crime)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36104 -0.19129  0.07939  0.27754  0.86843
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.86792    0.43153 -11.281  < 2e-16 ***
## log(prbarr)  -0.72397    0.11532  -6.278 1.39e-08 ***
## log(prbconv) -0.47251    0.08311  -5.686 1.80e-07 ***
## log(prbpris)  0.15967    0.20644   0.773    0.441
## log(avgsen)   0.07642    0.16347   0.467    0.641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.429 on 85 degrees of freedom
## Multiple R-squared:  0.4162, Adjusted R-squared:  0.3888
## F-statistic: 15.15 on 4 and 85 DF,  p-value: 2.171e-09
```

```
mod2 <- lm(crmrte ~ prbarr + prbconv + prbpris + avgsen + polpc, data=crime)
summary(mod2)
```

```
##
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + prbpris + avgsen + polpc,
##     data = crime)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.038338 -0.007831 -0.000843  0.006578  0.039503
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0551317  0.0094088   5.860 8.80e-08 ***
## prbarr      -0.0896980  0.0112238  -7.992 6.30e-12 ***
## prbconv     -0.0272321  0.0040173  -6.779 1.57e-09 ***
## prbpris      0.0121931  0.0173230   0.704    0.483
## avgsen      -0.0003331  0.0005662  -0.588    0.558
## polpc       10.5869888  1.7684871   5.986 5.11e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01304 on 84 degrees of freedom
## Multiple R-squared:   0.55,  Adjusted R-squared:  0.5232
## F-statistic: 20.53 on 5 and 84 DF,  p-value: 2.447e-13
```

```
mod2_log <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgsen) + log(polpc), data=
summary(mod2_log)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##     log(avgsen) + log(polpc), data = crime)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.40193 -0.22316  0.07635  0.25674  0.69986
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.70812    1.05253  -1.623  0.10836
## log(prbarr)  -0.73391    0.10934  -6.712 2.12e-09 ***
## log(prbconv) -0.43409    0.07965  -5.450 4.94e-07 ***
## log(prbpris)  0.13076    0.19587   0.668  0.50623
## log(avgsen)  -0.14564    0.16927  -0.860  0.39201
## log(polpc)    0.41412    0.12710   3.258  0.00162 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4066 on 84 degrees of freedom
## Multiple R-squared:  0.4817, Adjusted R-squared:  0.4509
## F-statistic: 15.62 on 5 and 84 DF,  p-value: 7.621e-11
```

I disagree with the inclusion of police percentage in this regression. While all else equal, we may expect more police to result in a deterrent, we can only assess this against an unobservable counterfactual - how the same location would have been impacted had more police or less police been assigned there. This would be possible to address experimentally, but not observationally. The problem is that places with increasing crime rates are subsequently likely to increase police presence, while those with decreasing crime rates are likely to decrease police presence (for budgetary reasons).

```
mod3 <- lm(crmrte ~ prbarr + prbconv + prbpris + avgsen + density, data=crime)
summary(mod3)
```

```
##
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + prbpris + avgsen + density,
##     data = crime)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.019774 -0.008035 -0.002675  0.005583  0.043581
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0347037  0.0084464   4.109 9.21e-05 ***
## prbarr      -0.0338981  0.0099540  -3.405 0.001015 **
## prbconv     -0.0148290  0.0037470  -3.958 0.000158 ***
## prbpris      0.0055122  0.0157162   0.351 0.726668
## avgsen       0.0004454  0.0004666   0.954 0.342582
## density      0.0072471  0.0009143   7.926 8.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01178 on 84 degrees of freedom
## Multiple R-squared:  0.6327, Adjusted R-squared:  0.6108
## F-statistic: 28.94 on 5 and 84 DF,  p-value: < 2.2e-16
```

```
mod3_log <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgsen) + log(density), da
summary(mod3_log)
```

```
##
```

```
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##     log(avgsen) + log(density), data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.17789 -0.21695  0.06725  0.25095  0.80522
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.57783    0.40718 -11.243  < 2e-16 ***
## log(prbarr)  -0.53503    0.11764  -4.548 1.81e-05 ***
## log(prbconv) -0.41601    0.07845  -5.303 9.07e-07 ***
## log(prbpris) -0.23672    0.21732  -1.089 0.279155
## log(avgsen)  -0.08037    0.15695  -0.512 0.609946
## log(density)  0.14780    0.03835   3.854 0.000227 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3978 on 84 degrees of freedom
## Multiple R-squared:  0.5039, Adjusted R-squared:  0.4744
## F-statistic: 17.07 on 5 and 84 DF,  p-value: 1.292e-11
```

I also disagree with this simplistic inclusion of density (if our goal is causation, not prediction). All else equal - density means more eyes looking around - which should mean less propensity to commit crime. There are likely other variables which better explain what many lazily associate with density. For example, thresholds of inequality, where extreme wealth exists right next to extreme poverty are more common in more dense locations. Additionally, more dense locations sometimes have less average social capital (perhaps again because of close proximity class differences) than less dense places. These are good instances of omitted variable bias. Since our goal is causation rather than prediction, we should be cautious in including density as an explanatory variable.

```
mod4 <- lm(crmrte ~ prbarr + prbconv + prbpris + avgsen + pctymle, data=crime)
summary(mod4)
```

```
##
## Call:
## lm(formula = crmrte ~ prbarr + prbconv + prbpris + avgsen + pctymle,
##     data = crime)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.031869 -0.008483 -0.002073  0.006987  0.053989
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0333499  0.0129958   2.566   0.0121 *
## prbarr      -0.0583518  0.0123644  -4.719 9.35e-06 ***
## prbconv     -0.0219883  0.0047617  -4.618 1.38e-05 ***
## prbpris      0.0231725  0.0203451   1.139   0.2580
## avgsen       0.0010613  0.0006008   1.767   0.0809 .
## pctymle      0.1154660  0.0725020   1.593   0.1150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.01535 on 84 degrees of freedom
## Multiple R-squared:  0.3768, Adjusted R-squared:  0.3397
## F-statistic: 10.16 on 5 and 84 DF,  p-value: 1.24e-07
```

```
mod4_log <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) + log(avgsen) + log(pctymle), da
summary(mod4_log)
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(prbpris) +
##     log(avgsen) + log(pctymle), data = crime)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3599 -0.1852  0.0841  0.2844  0.8779
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.35261    0.86966  -5.005 3.03e-06 ***
## log(prbarr)  -0.69480    0.12331  -5.635 2.29e-07 ***
## log(prbconv) -0.45446    0.08746  -5.196 1.40e-06 ***
## log(prbpris)  0.16249    0.20714   0.784    0.435
## log(avgsen)   0.06348    0.16508   0.385    0.702
## log(pctymle)  0.17282    0.25296   0.683    0.496
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4304 on 84 degrees of freedom
## Multiple R-squared:  0.4195, Adjusted R-squared:  0.3849
## F-statistic: 12.14 on 5 and 84 DF,  p-value: 7.347e-09
```

The inclusion of the percent of the population that is young and male, as expected, has a notable positive relationship with crime rate. The variable is only marginally significant though. I would argue that, since we are interested in causation, particularly in differentiating the effect of certainty of punishment from severity of punishment, we should include this variable so as to make the coefficients for certainty and severity of punishment less biased.

The logged version of this regression is not anywhere close to significant for the log(pctymle) variable.

Interesting things to note after examining these regression approaches:

1) The certainty of punishment is estimated much more consistently than any other variable. In every regression run, these variables have a clear, significant negative relationship with crime rate. This is a good basis upon which we can make policy recommendations -Our policy recommendations focus on increasing the certainty of punishment in areas with low certainty of punishment now. Police departments need quality information in order to increase the certainty of punishment, so improving relationships with local people is a key recommendation. Additionally, locations with a low conviction rate could invest in more prosecutorial staff/better prosecutorial staff.
2) The severity of punishment seems to be much less consistent, with the sign flipping depending on the regression specification. In contrast to some people's expectations, most regressions show that the probability of going to prison is associated with a higher crime rate, rather than a lower one. One way of thinking about this is that prisons have become cultivators of more organized crime. Another way to think about it is prisons currently do little to change the person put in them for the better. Rather, prisoners associate all day every day with other prisoners - perhaps exchanging bad habits and temperaments and building social connections that may persist in and outside of prison.
3) Police presence and density are both useful predictors, though we believe that they actually obscure the causal relationship we desire to study. Police presence tends to increase along with crime and decrease

as crime goes away. In other words, the direction of causality is likely reversed from what we have specified here. Secondly, density is a useful predictor, but we feel it does not contain causal information about crime rates. Rather density obscures more relevant but unobserved relationships with other factors such as proximal inequality.

Steps for evaluating variables

Leverage (and Influence if required)
Goodness-of-Fit : AIC
Omitted variable bias
MSE
E[theta hat] = theta

```
crime$urban + crime$west + crime$central
```

```
##  [1] 1 1 1 1 1 1 0 0 0 0 2 1 1 1 1 1 1 1 0 1 0 0 1 0 0 1 1 0 2 0 2 1 2 1 0
## [36] 2 0 0 1 1 0 0 1 1 0 1 0 1 1 1 1 0 2 1 1 0 1 0 0 1 0 0 0 0 1 0 1 1 1 0
## [71] 1 1 1 0 0 1 1 1 1 1 1 1 2 1 0 1 0 1 0 1
```

```
crime$urban + crime$west
```

```
##  [1] 0 0 1 0 1 1 0 0 0 0 2 1 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 0 1 0 0
## [36] 1 0 0 1 1 0 0 0 1 0 0 0 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0
## [71] 0 0 1 0 0 0 0 1 1 1 0 0 1 0 0 1 0 1 0 1
```

```
crime$urban + crime$central
```

```
##  [1] 1 1 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 1 0 0 1 0 0 1 0 0 1 1 0 2 0 2 1 1 1 0
## [36] 2 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 2 1 1 0 1 0 0 1 0 0 0 0 1 0 0 1 1 0
## [71] 1 1 0 0 0 1 1 0 0 0 1 1 2 1 0 0 0 0 0 0
```

```
crime$west + crime$central
```

```
##  [1] 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1 0 1 0 0 0 0 0 1 1 0 1 0 1 1 2 1 0
## [36] 1 0 0 1 1 0 0 1 1 0 1 0 1 1 1 1 0 1 1 1 0 0 0 0 1 0 0 0 0 1 0 1 1 1 0
## [71] 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1
```

Conclusion

From the assignment "Since you are restricted to ordinary least squares regression, omitted variables will be a major obstacle to your estimates. You should aim for causal estimates, while clearly explaining how you think omitted variables may affect your conclusions."

Apendix A: Codebook

Table 1: Crime Data Codebook

Variable	Label	Notes
county	county identifier	
year	1987	
crmrte	crimes committed per person	
prbarr	'probability' of arrest	County 115 has a value of 1.09, which is not a possible probability.
prbconv	'probability' of conviction	There are 10 observations greater than 1, which is not a possible probability.
prbpris	'probability' of prison sentence	
avgsen	avg. sentence, days	
polpc	police per capita	
density	people per sq. mile	
taxpc	tax revenue per capita	
west	=1 if in western N.C.	
central	=1 if in central N.C.	
urban	=1 if in SMSA	
pctmin80	perc. Minority, 1980	
wcon	weekly wage, construction	
wtuc	wkly wge, trns, util, commun	
wtrd	wkly wge whlesle, retail, trade	
wfir	wkly wge, fin, ins, real est	
wser	wkly wge, service industry	
wmfg	wkly wge, manufacturing	
wfed	wkly wge fed employees	
wsta	wkly wge state employees	
wloc	wkly wge local gov emps	
mix	offense mix: face-to-face/other	
pctymle	percent young male	