lab3: Reducing Crime

Thomas Drage, Venkatesh Nagapudi, Miguel Jamie

November 20, 2018

1. Introduction

This statistical investigation is aimed at understanding the determinants of crime in order to generate policy suggestions that are applicable to the local government. The study is based upon development of causal models for crime rate, based on county level demographic and judicial data for 1987. We have identified factors which modify the rate and extended this to the development of policy proposals for a new government.

What we are graded on:

Introduction. As you understand it, what is the motivation for this team's report? Does the introduction as written make the motivation easy to understand? Is the analysis well-motivated? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.

2. Review of Source Data

```
rm(list = ls())
crime_data = read.csv("crime_v2.csv")
objects(crime_data)
```

```
##  [1] "avgsen"   "central"  "county"   "crmrte"   "density"  "mix"
##  [7] "pctmin80" "pctymle"  "polpc"    "prbarr"   "prbconv"  "prbpris"
## [13] "taxpc"    "urban"    "wcon"     "west"     "wfed"     "wfir"
## [19] "wloc"     "wmfg"     "wser"     "wsta"     "wtrd"     "wtuc"
## [25] "year"
```

Finding out number of observations

```
str(crime_data)
```

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
```

```
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

There are 97 of them.

Data Cleansing

1. Removing NA in some cases

```
crime_data_corr = na.omit(crime_data)
```

2. Some values are coded as levels: prbconv - need to fix

```
crime_data_corr$prbconv_fix = as.numeric(as.character(crime_data_corr$prbconv))
summary(crime_data_corr$prbconv_fix)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06838 0.34541 0.45283 0.55128 0.58886 2.12121
```

3. Probability values > 1 in some cases. There are 11 such values. Perhaps we have to leave these rows out.

```
sum(crime_data_corr$prbarr > 1)
```

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

```
sum(crime_data_corr$prbconv_fix > 1)
```

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

```
sum(crime_data_corr$prbpris > 1)
```

```
## [1] 0
```

Eliminate the above points from the data set

```
good_prob_cond =
   !((crime_data_corr$prbarr > 1) |
   (crime_data_corr$prbconv_fix > 1) |
   (crime_data_corr$prbpris > 1))
crime_data_corr2 = subset (crime_data_corr, good_prob_cond)
str(crime_data_corr2)
```

```
## 'data.frame':    81 obs. of  26 variables:
##  $ county     : int  1 5 7 9 11 13 15 17 21 23 ...
##  $ year       : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte     : num  0.0356 0.013 0.0268 0.0106 0.0146 ...
##  $ prbarr     : num  0.298 0.444 0.365 0.518 0.525 ...
##  $ prbconv    : Factor w/ 92 levels "","`","0.068376102",..: 63 13 62 52 3 59 78 42 23 37 ...
##  $ prbpris    : num  0.436 0.6 0.435 0.443 0.5 ...
##  $ avgsen     : num  6.71 6.76 7.14 8.22 13 ...
```

```
##  $ polpc      : num  0.00183 0.00123 0.00153 0.00086 0.00288 ...
##  $ density    : num  2.423 0.413 0.492 0.547 0.611 ...
##  $ taxpc      : num  31 34.8 42.9 28.1 35.2 ...
##  $ west       : int  0 1 0 1 1 0 0 0 1 1 ...
##  $ central    : int  1 0 1 0 0 0 0 0 0 0 ...
##  $ urban      : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ pctmin80   : num  20.22 3.16 47.92 1.8 1.54 ...
##  $ wcon       : num  281 227 375 292 250 ...
##  $ wtuc       : num  409 372 398 377 401 ...
##  $ wtrd       : num  221 229 191 207 188 ...
##  $ wfir       : num  453 306 281 289 259 ...
##  $ wser       : num  274 210 257 215 237 ...
##  $ wmfg       : num  335 238 282 291 259 ...
##  $ wfed       : num  478 359 412 377 391 ...
##  $ wsta       : num  292 332 328 367 326 ...
##  $ wloc       : num  312 281 299 343 275 ...
##  $ mix        : num  0.0802 0.4651 0.2736 0.0601 0.3195 ...
##  $ pctymle    : num  0.0779 0.0721 0.0735 0.0707 0.0989 ...
##  $ prbconv_fix: num  0.5276 0.2679 0.5254 0.4766 0.0684 ...
```

4. There is a duplicate entry for county #193, which we will also remove from the data set.

```
crime_data_corr2[crime_data_corr2$county == 193, 1:6]
```

```
##    county year    crmrte   prbarr     prbconv  prbpris
## 88    193   87 0.0235277 0.266055 0.588859022 0.423423
## 89    193   87 0.0235277 0.266055 0.588859022 0.423423
```

```
crime_data_corr3 = crime_data_corr2[!duplicated(crime_data_corr2), ]
```

Now there are only 80 observerations which is less than 100. So we do need to be careful our assumptions around CLM for coefficients being normal.

Once data cleaning is complete, creating a working copy of our data

```
crime_data_clean = crime_data_corr3
```

3. Identification of Key Variables

Dependent Variable

The crime rate ("crmrte") is the key dependent variable in this study and represents the number of crimes committed per person in the each county.

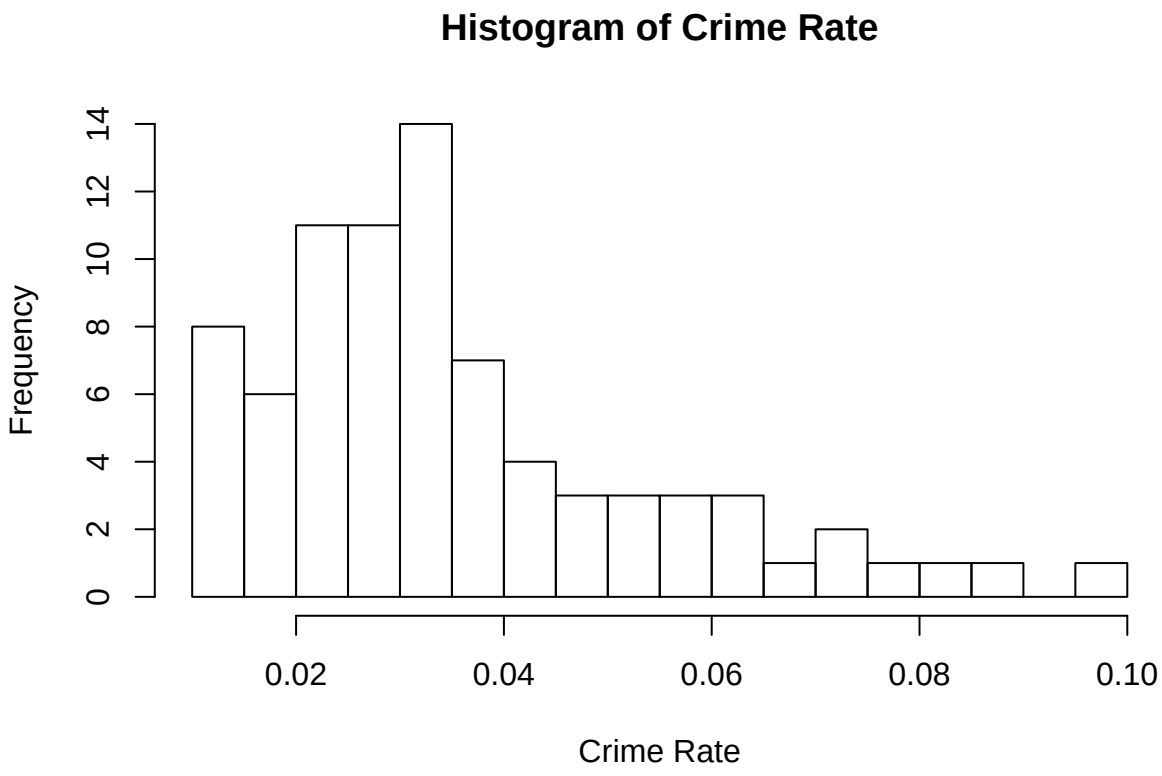
Summarizing the variable we note a small range of fractional values, centred on a mean of approximately 3.5 crimes per hundred people in the year period.

```
summary(crime_data_clean$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02336 0.03051 0.03551 0.04385 0.09897
```

The distribution of crime rate is somewhat left-skewed in this dataset but sufficient data is available for modelling.

```r
hist(crime_data_clean$crmrte, breaks = 30,
     main = 'Histogram of Crime Rate',
     xlab = 'Crime Rate' )
```



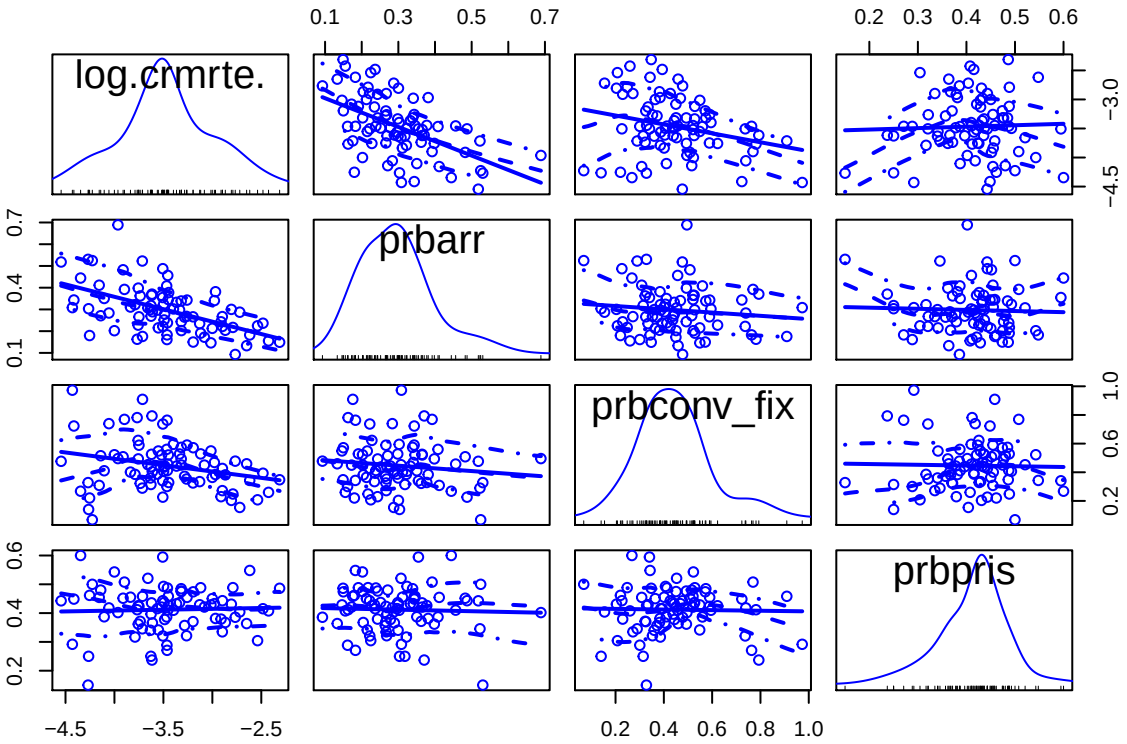
Independent Variables - Judicial

1. Probablity of Arrest ("prbarr")
2. Probability of Conviction ("prbconv")
3. Probability of Going to Prison ("prbpris")

The assumption here is that crime rate will be lower if the probability of getting arrested, convicted or going to prison is higher. Crimes happen if criminals believe that they can get away with performing crimal acts since the probability of getting punished is lower.

Let's look at the scatterplot matrix for a relationship with crmrte

```r
scatterplotMatrix(~ log(crmrte) + prbarr +  prbconv_fix + prbpris, data=crime_data_clean)
```



As we can see, the log(crimrte) seems to be negatively correlated with prbarr and prbconv_fix which seems to be intuitive. There is perhaps a positive correlation to prbpris, which is not very intuitive.

But on the flipside, what really motivates or prevents crime? Should we have some of those variables in here as primarly independent variables?

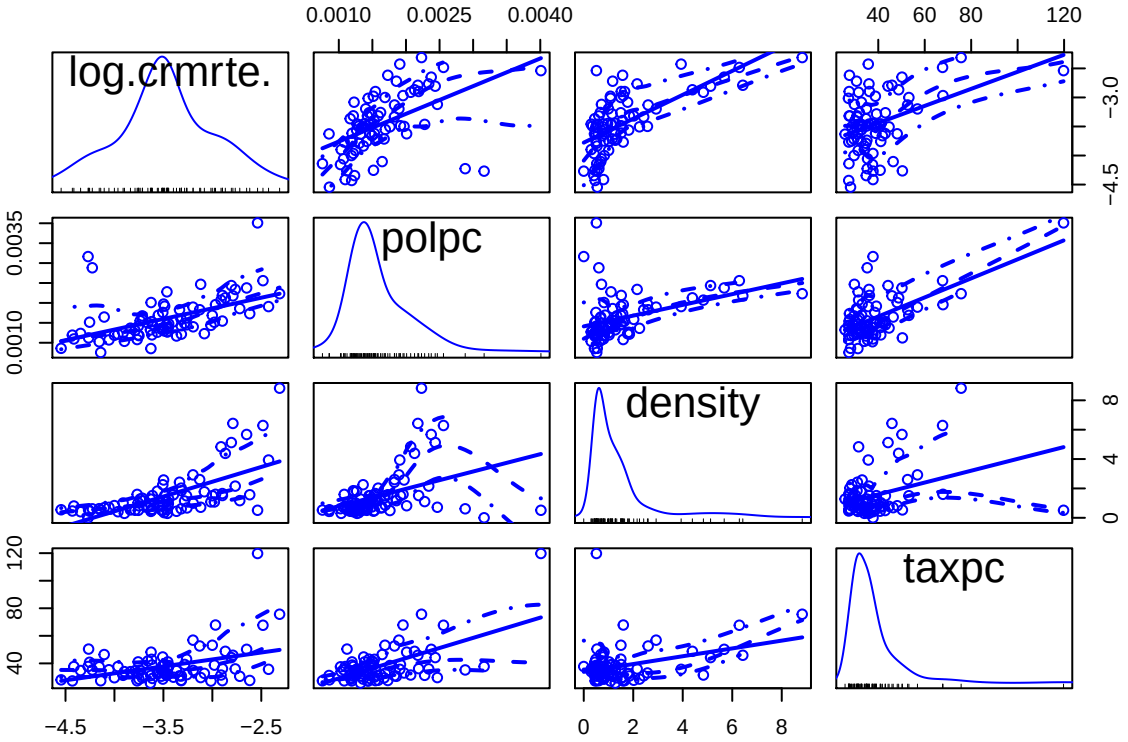
Idependent Variables - Demographic

Should some of these be key independent variables? 1. Police per capita ("polpc") 2. Density ("density") 3. Tax revenue per capita ("taxpc") 4. Percentage of Young males ("pctymle") 5. Percentage of minorities ("pctmin80") 6. Average sentence ("avgsen")

Crime rate likely depends on the deterrents to crime: police protection But it is likely crime is high if the county is poor or has young males/minorities (ex: Oakland?) but not when the county is rich (there are people to rob, but then there will be better protection as well like security alarms etc).

Performing a couple of different scatterplots

```
scatterplotMatrix(~ log(crmrte) + polpc +  density + taxpc, data=crime_data_clean)
```



Crime Rate seems to be positively correlated to the "Police per capita". If we consider police staffing as a lagging indicator, this is intuitive: where Crime Rate is high, more police officers will be deployed. This would be an inverse causal relationship.

Looking at population density, there is a positive correlation between crime and density. This seems intuitive. However, the density distribution is not very normal, and might need a transformation.

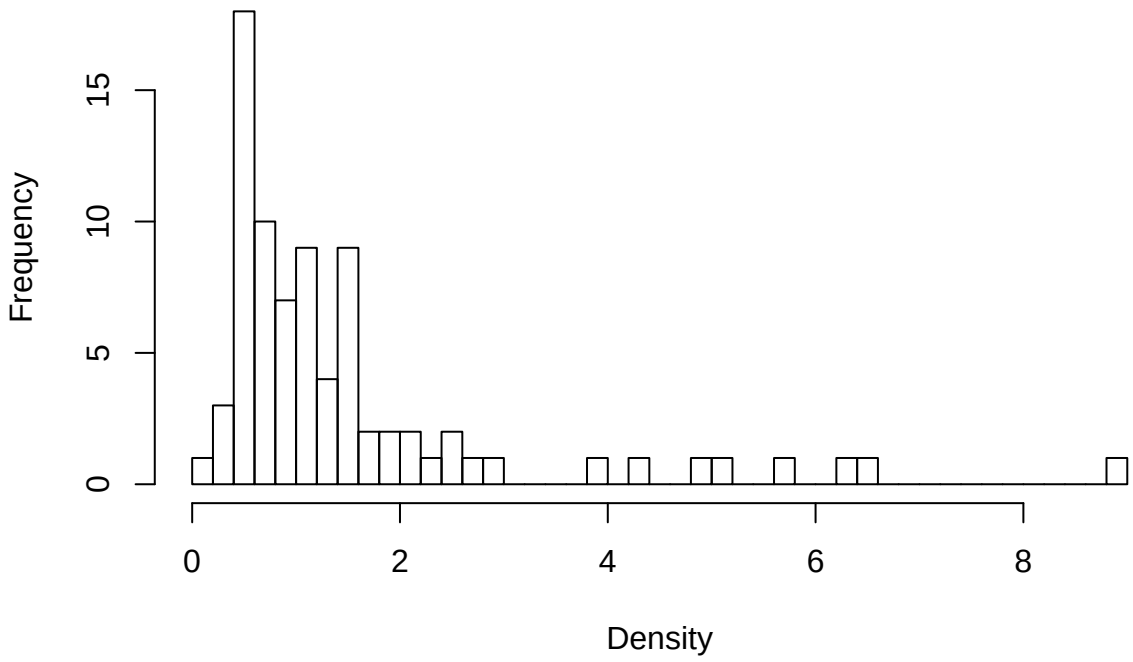
```
summary(crime_data_clean$density)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00002 0.55971 1.01575 1.51705 1.59508 8.82765
```

```
cor(crime_data_clean$crmrte,crime_data_clean$density)
```

```
## [1] 0.7235116
```

```
hist(crime_data_clean$density, breaks = 50,
     main = 'Histogram of Density',
     xlab = 'Density' )
```



The Taxpc is a proxy for how rich a county is. It is likely that the higher the tax paid, the more likely that the people are, on average, richer. On one hand, richer counties might be a more attractive target for property crime. On the other hand, people in this counties have less of an economic incentive to commit crime, and are likely to have better security measures than less rich counties.

```
summary(crime_data_clean$taxpc)
```

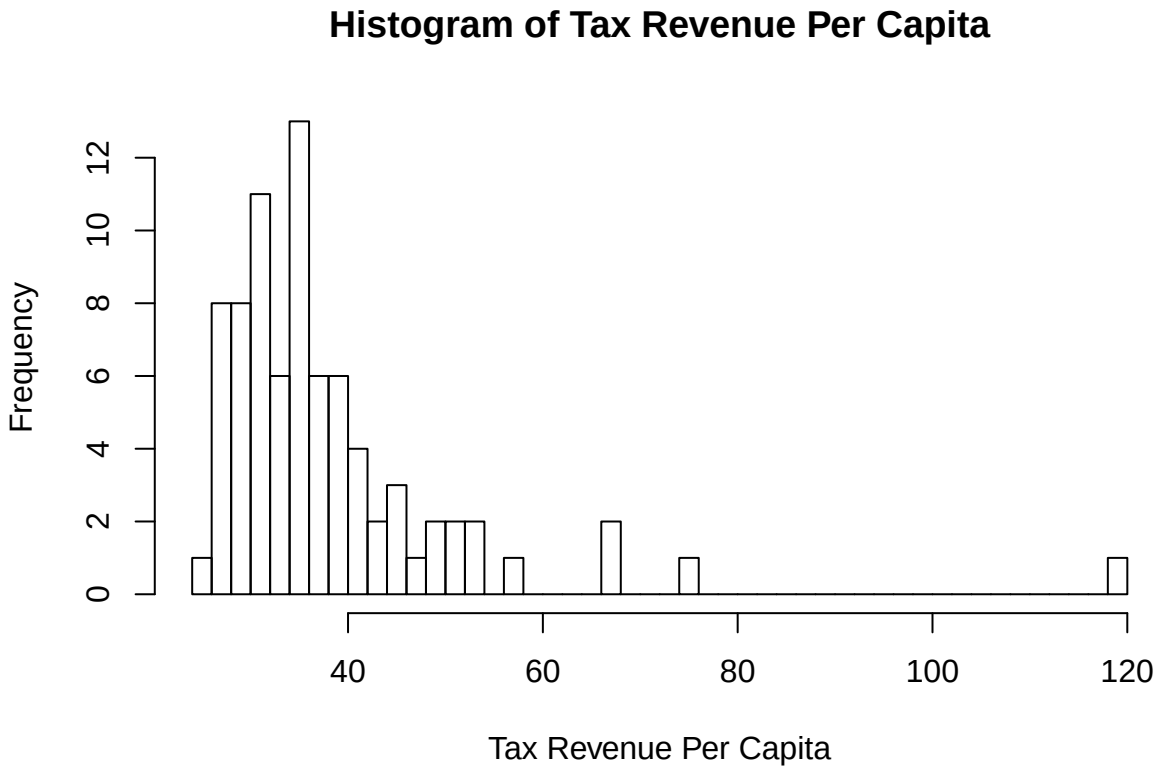
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   25.69   30.96   34.92   38.16   40.87  119.76
```

```
cor(crime_data_clean$crmrte, crime_data_clean$taxpc)
```

```
## [1] 0.4772213
```

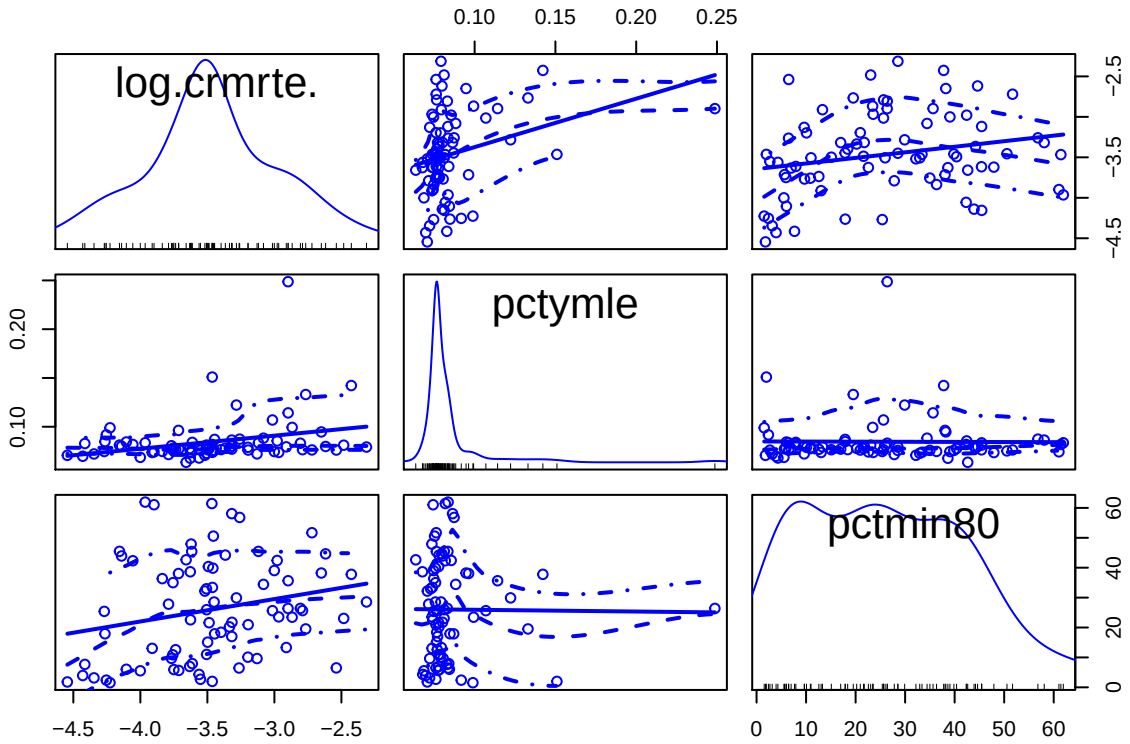
Look at the correlation, we see a positive correlation between taxpc and crime rate. However, the distribution of taxpc is not very optimal and we do need to watch out for outliers creating a lot of leverage and influence.

```
hist(crime_data_clean$taxpc, breaks = 50,
     main = 'Histogram of Tax Revenue Per Capita',
     xlab = 'Tax Revenue Per Capita' )
```



Examining the relationship between pctymle and pctmin80 with crmrte:

```
scatterplotMatrix(~ log(crmrte) + pctymle +  pctmin80, data=crime_data_clean)
```



The crime rate is higher in places with more % of young males. This seems somewhat likely. The crime rate is higher generally when minority % is higher. However, both variables seem to have non-ideal distributions.

(NOTE(miguel): seemscorrelation is not very high, but it might still be important, IMO.)

Looking at the correlation between the variables:

```
summary(crime_data_clean$pctymle)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06356 0.07510 0.07791 0.08463 0.08367 0.24871
```

```
cor(crime_data_clean$crmrte,crime_data_clean$pctymle)
```

```
## [1] 0.2909047
```

```
summary(crime_data_clean$pctmin80)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.541  10.674  25.510  26.022  38.739  61.942
```

```
cor(crime_data_clean$crmrte,crime_data_clean$pctmin80)
```

```
## [1] 0.1738522
```

The correlation is not very high though.

Finally looking at avgsen,

```
summary(crime_data_clean$avgsen)
```

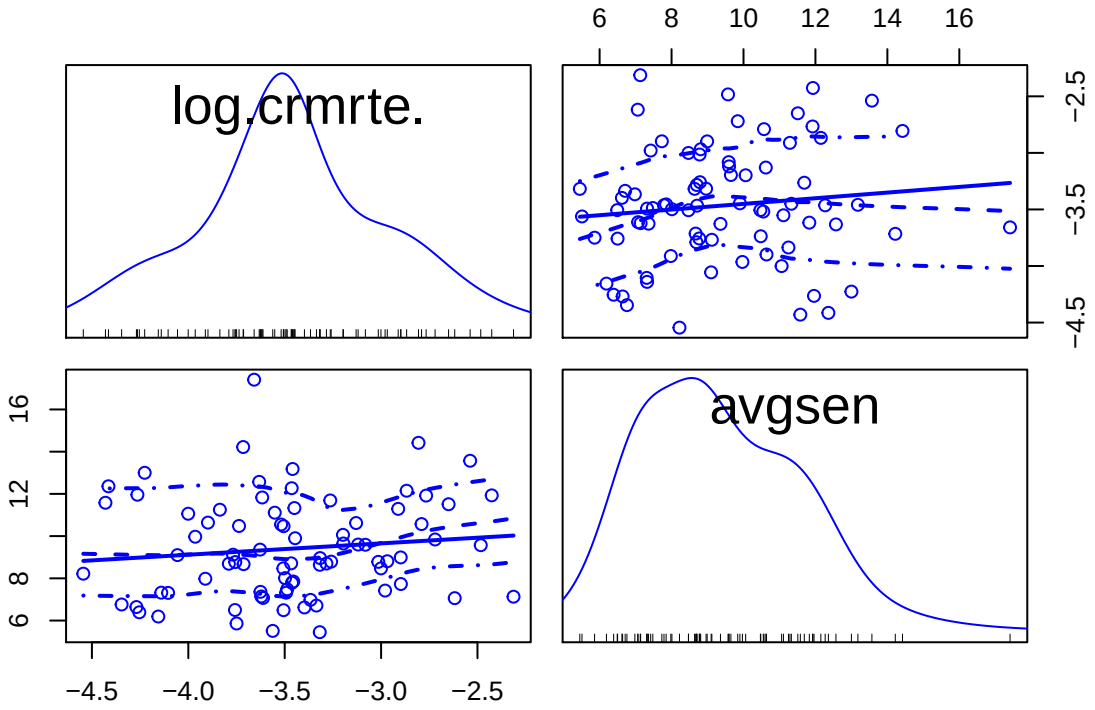
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.450   7.405   8.975   9.405  11.145  17.410
```

```
cor(crime_data_clean$crmrte,crime_data_clean$avgsen )
```

```
## [1] 0.1346185
```

There is a small correlation here. But it is unclear as to whether there will be a causal relationship and which way it would be directed.

```
scatterplotMatrix(~ log(crmrte) + avgsen, data=crime_data_clean)
```



3. Data Transformation

Not sure if we have transformations in this section or in the later models section:

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.

Some inputs from today's post-class session: 1. Use a log transformation on crime rate since the values are very small 2. Apply transformations in X variables and try to figure out if r.square improves or MSE goes down (this requires a model to be built though) 3. There was a discussion on Y-transformation which I didn't understand at all... not sure what that is... perhaps week 12 async has it? 4. If you apply Y-transformation, apply it universally (Prof said this: not sure what it means!)

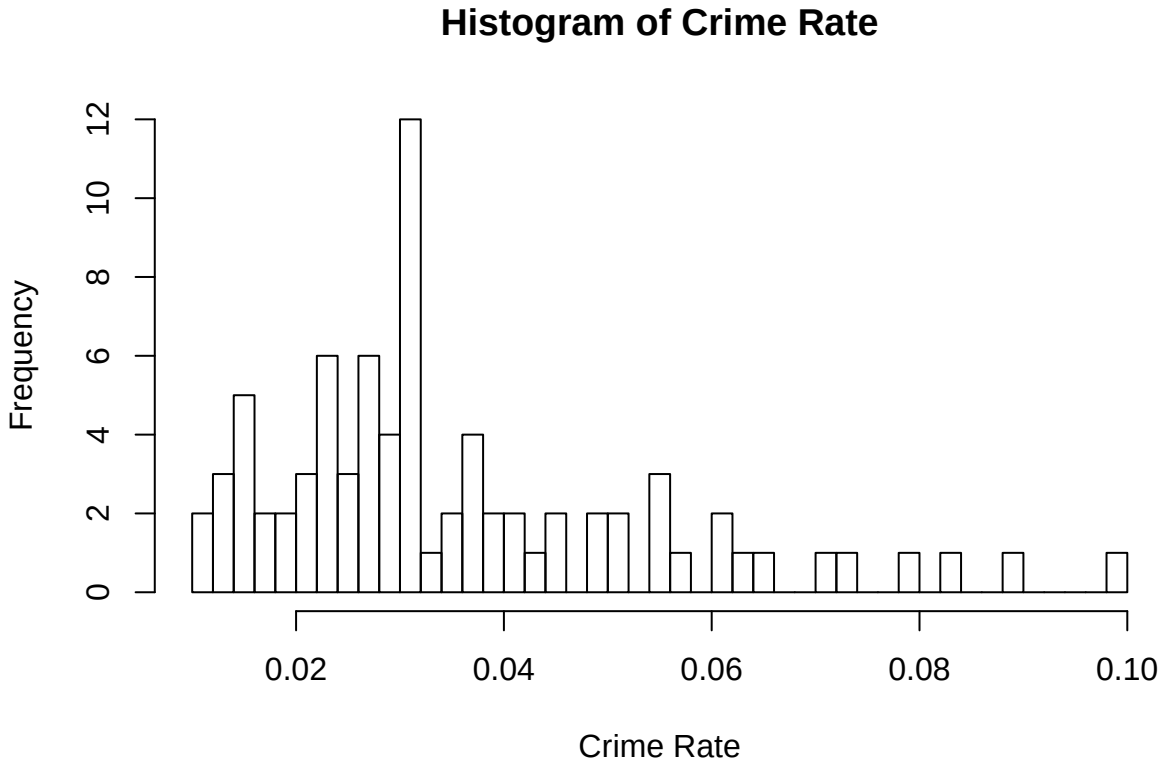
Crime Rate

As discussed in section 2, our main variable of interest, crime rate, is measured in a way that results in small variations between values, and a skewed distribution:

```
summary(crime_data_clean$crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01062 0.02336 0.03051 0.03551 0.04385 0.09897
```

```
hist(crime_data_clean$crmrte, breaks = 50,
     main = 'Histogram of Crime Rate',
     xlab = 'Crime Rate' )
```

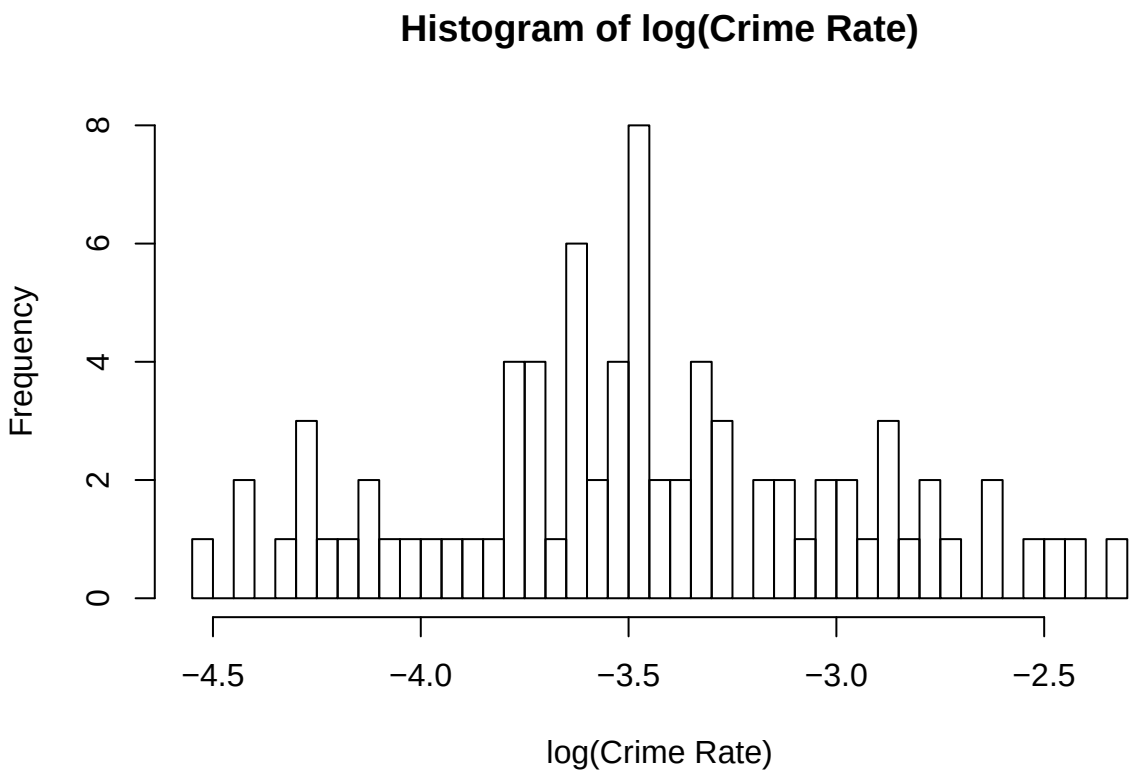


As a result, we will apply a log() transformation to the variable, which will address both issues. This transformation will change our interpretation as well, since the result will need to be interrpeted as a percentage change for Crime Rate..

```
crime_data_clean['log_crmrte'] = log(crime_data_clean$crmrte)
summary(crime_data_clean$log_crmrte)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -4.545  -3.757  -3.490  -3.466  -3.127  -2.313
```

```
hist(crime_data_clean$log_crmrte, breaks = 50,
     main = 'Histogram of log(Crime Rate)',
     xlab = 'log(Crime Rate)' )
```



Justice System Effects

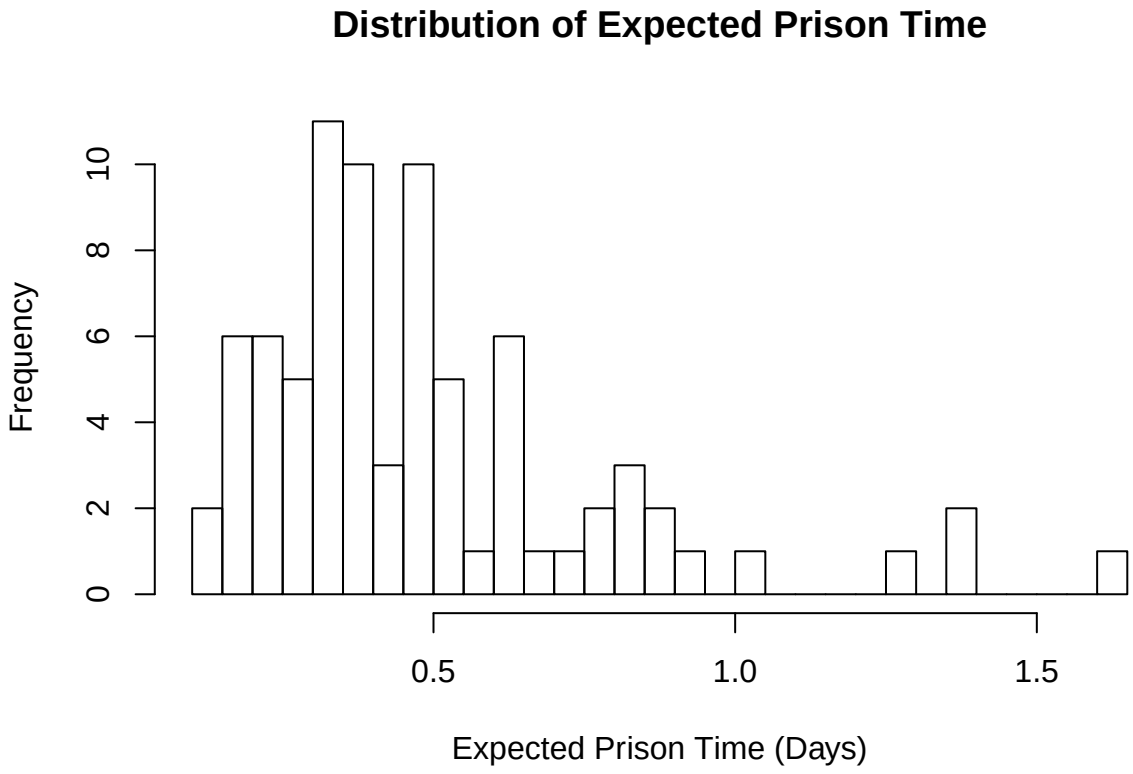
The data contains several variables related to the potential consequences for a person committing a crime. These are the probabilities of being arrested, convicted, sentenced to prison, and the average length of said sentece.

Instead of using the variables individually, we will condense them into one, which will incorporate the probabilities of each step as well as the sentence. This variable, which we will call "expected time in prison" or "exp_pris_time", will be obtained by multiplying each probability and the expected average sentence.

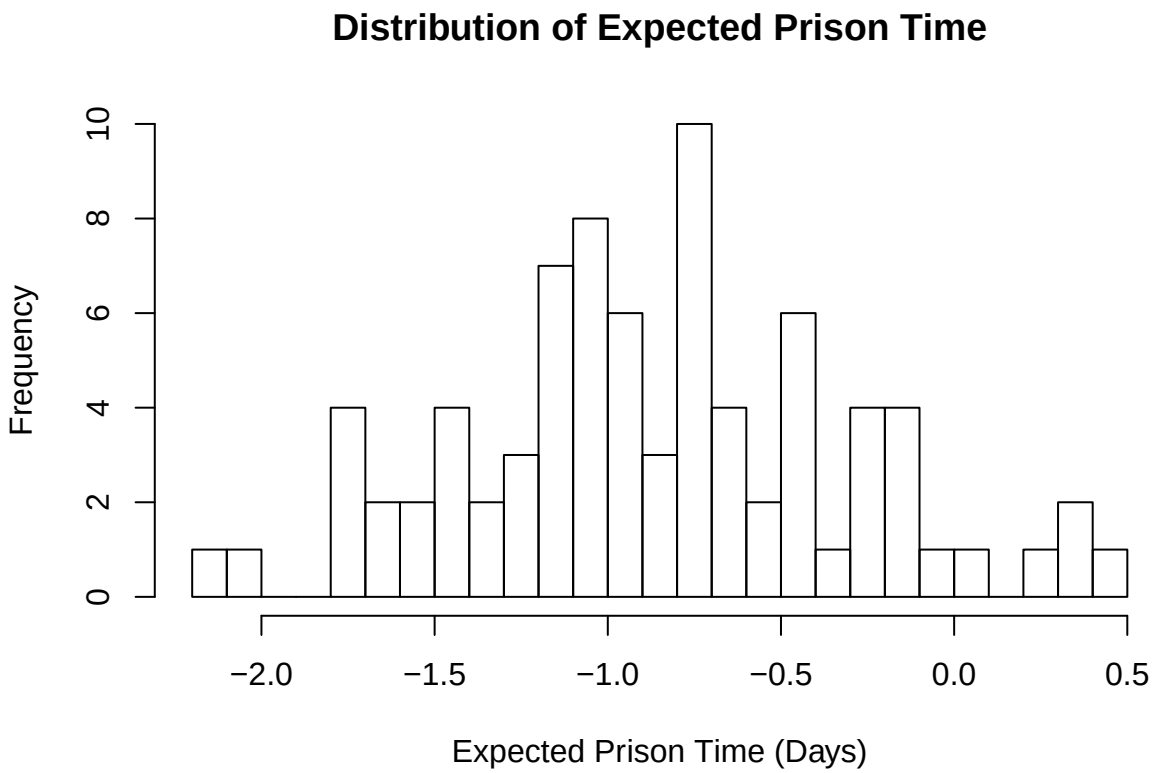
```
crime_data_clean["exp_pris_time"] = crime_data_clean$prbarr * crime_data_clean$prbconv_fix * crime_data_
```

The resulting variable is right-skewed, so will then take the log, which yields a more normal distribution, and use that variable going forward.

```
hist(crime_data_clean$exp_pris_time, breaks = 30,
     main = 'Distribution of Expected Prison Time',
     xlab = 'Expected Prison Time (Days)' )
```



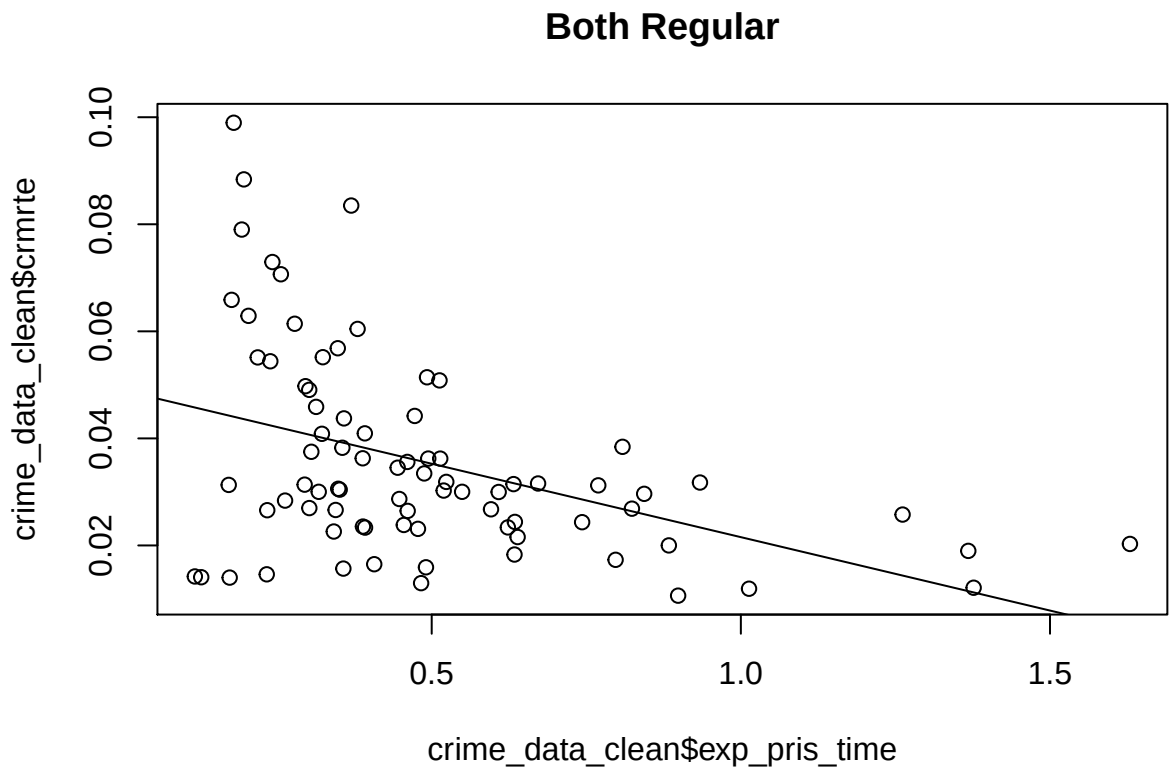

```
hist(log(crime_data_clean$exp_pris_time), breaks = 30,
     main = 'Distribution of Expected Prison Time',
     xlab = 'Expected Prison Time (Days)' )
```



```
crime_data_clean["log_exp_pris_time"] = log(crime_data_clean$exp_pris_time)
```

By using both log variables (Crime Rate and Expected Prison Time), we get a less heteroskedastic distribution between our variables, as illustrated by the plots below.

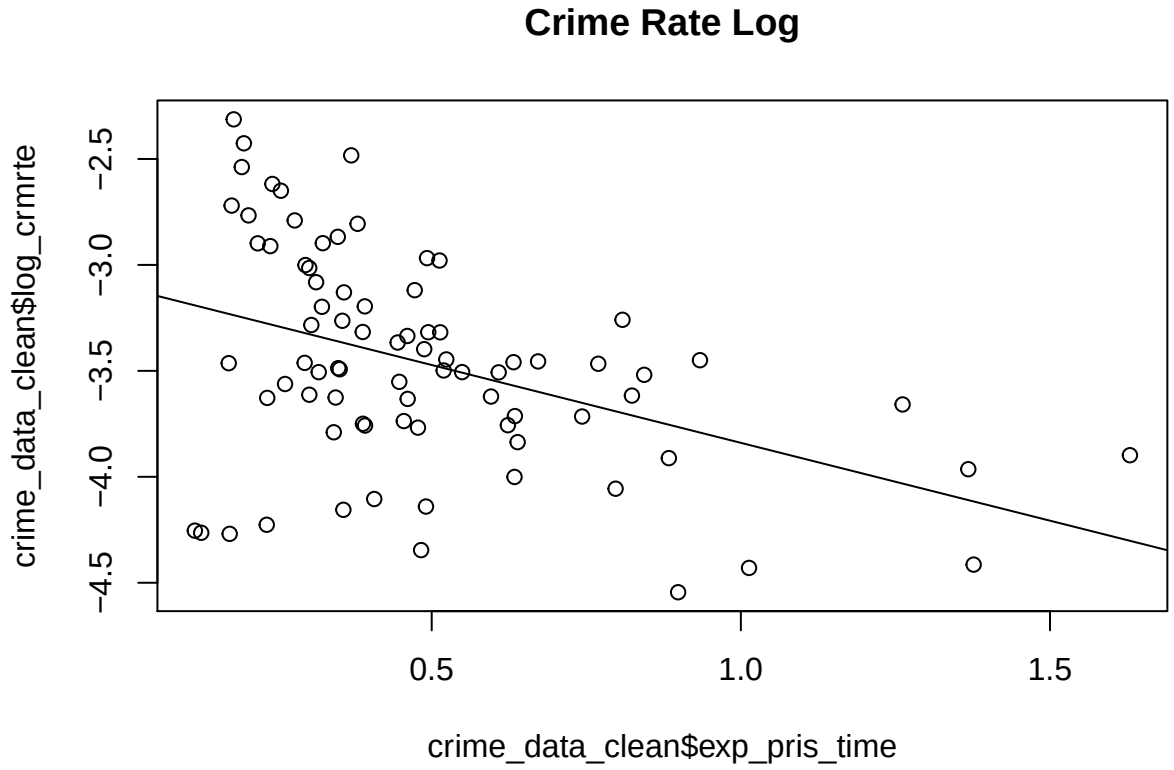
```
lm1 = lm(crime_data_clean$crmrte ~ crime_data_clean$exp_pris_time)
plot(crime_data_clean$exp_pris_time, crime_data_clean$crmrte,
     main = 'Both Regular')
abline(lm1)
```




```
summary(lm1)$adj.r.squared
```

```
## [1] 0.1736396
```

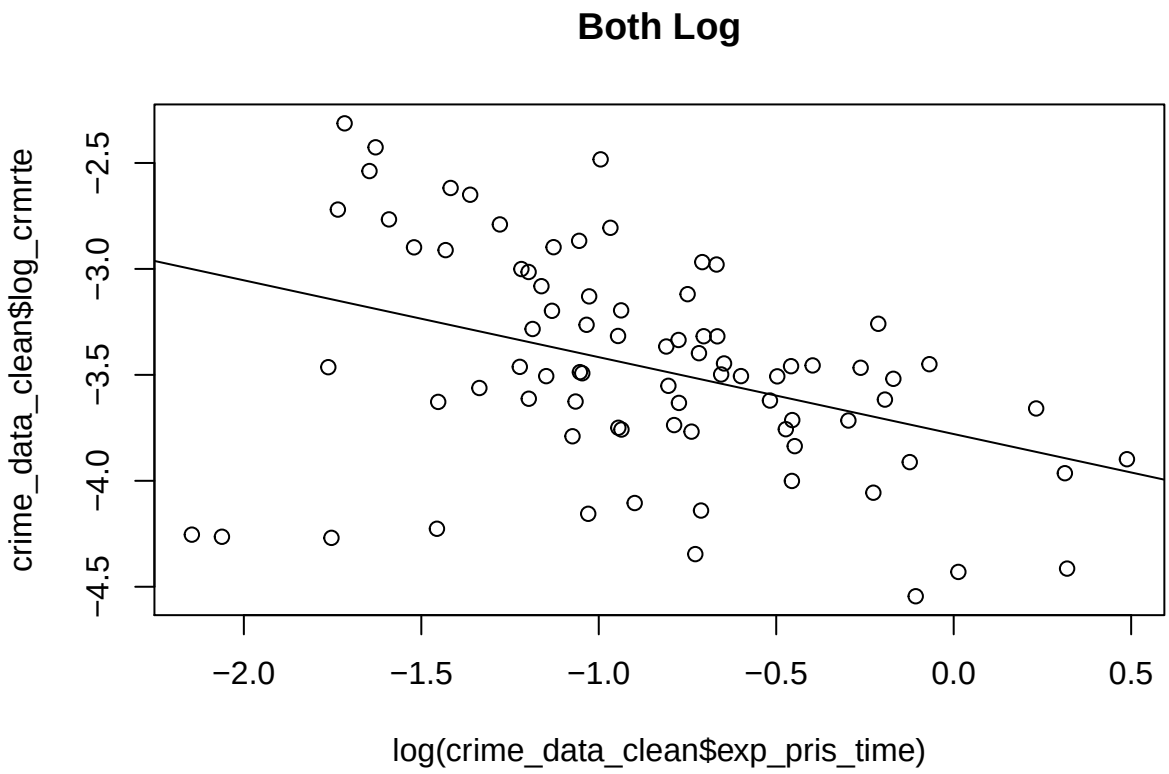
```
lm2 = lm(crime_data_clean$log_crmrte ~ crime_data_clean$exp_pris_time)
plot(crime_data_clean$exp_pris_time, crime_data_clean$log_crmrte,
     main = 'Crime Rate Log')
abline(lm2)
```




```
summary(lm2)$adj.r.squared
```

```
## [1] 0.1729452
```

```
lm3 = lm(crime_data_clean$log_crmrte ~ log(crime_data_clean$exp_pris_time))
plot(log(crime_data_clean$exp_pris_time), crime_data_clean$log_crmrte,
     main = 'Both Log')
abline(lm3)
```




```
summary(lm3)$adj.r.squared
```

```
## [1] 0.1442286
```

4. Regression Modelling

Model 1

One model with only the explanatory variables of key interest (possibly transformed, as determined by your EDA), and no other covariates Things to do: - Get list of key explanatory variables - Should we find the correlation between them? - Get linear model going for crimerate against these variables - Get AIC, r.squared and other key elements of MLR

```
model1 = lm(crime_data_clean$crmrte ~ crime_data_clean$prbarr + crime_data_clean$prbconv_fix + crime_da
model3 = lm(crime_data_clean$crmrte ~ crime_data_clean$prbarr + crime_data_clean$prbconv_fix + crime_da
stargazer(model1, model3,  type = "latex", float=FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 26, 2018 - 00:12:52

	Dependent variable:	
	crmrte	
	(1)	(2)
prbarr	−0.095***	−0.058***
	(0.016)	(0.010)
prbconv_fix	−0.040***	−0.010
	(0.010)	(0.007)
prbpris	0.007	0.009
	(0.022)	(0.013)
avgsen	0.0004	−0.0003
	(0.001)	(0.0004)
polpc		8.038***
		(2.450)
density		0.005***
		(0.001)
taxpc		0.0003***
		(0.0001)
pctmin80		0.0004***
		(0.0001)
pctymle		0.118***
		(0.044)
Constant	0.075***	0.005
	(0.015)	(0.012)
Observations	80	80
R^2	0.398	0.828
Adjusted R^2	0.366	0.806
Residual Std. Error	0.015 (df = 75)	0.008 (df = 70)
F Statistic	12.385*** (df = 4; 75)	37.454*** (df = 9; 70)
Note:		*p<0.1; **p<0.05; ***p<0.01

```
AIC(model1, model3)
```

```
##        df       AIC
## model1  6 -437.3420
## model3 11 -527.6143
```

Model 2

One model that includes key explanatory variables and only covariates that you believe increase the accuracy of your results without introducing substantial bias (for example, you should not include outcome variables that will absorb some of the causal effect you are interested in). This model should strike a balance between

accuracy and parsimony and reflect your best understanding of the determinants of crime.

Things to do: - Get list of key secondary explanatory variables - Find correlation with key explanatory variables (is this easy?) - Get linear model going for crimerate against these variables - Get AIC, r.squared and other key elements of MLR

Model 3

One model that includes the previous covariates, and most, if not all, other covariates. A key purpose of this model is to demonstrate the robustness of your results to model specification.

Things to do: - This is the kitchensink where you throw everything in - Get linear model going for crimerate against these variables - Get AIC, r.squared and other key elements of MLR

Guided by your background knowledge and your EDA, other specifications may make sense. You are trying to choose points that encircle the space of reasonable modeling choices, to give an overall understanding of how these choices impact results.

What we learned from class today: 1. We need to apply transformations for sure 2. We need to perform AIC analysis as we add more models. If AIC is worse, then the added X value is not very useful 3. Another way of figuring out if a model is good is to look at the overall MSE and r.square 4. Of the above 4 models, model1 is basic, model2 is a bit more elaborate, model3 is the "kitchen sink". Model2 is supposed to be the optimized one wrt what X variables to use 5. If multi-collinearity is violated, the model will blow up and not converge 6. Apparently a lot of the model related info is in Week 12 async 7. Watch out for outliers in X variables... something about not having too much of a concentration towards the ends... 8. Variance of Beta is dependent on 1) sigmaˆ2 2)R-square and 3)SST. Async 12 has more material on this

5. Discussion - Model Specification & Omitted Variables

Some inputs from class: What we need to discuss is what columns were omitted that could help with getting a better model...

It is likely that crime rate will be heavily influenced by the following omitted variables: 1. Demographics: There is very little information on demographics other than pctmin80 which is based on dated information about minorities. It could be useful to get a bigger idea on the demographics of the county population. 2. Education level: The higher the education level, the lower the crime rate 3. Wages: The more affluent neighborhoods will tend to have lesser crime. This is somewhat reflected by the tax revenues per capita 4. Private Security: The higher the private security level, the lower the crime rate 5. Number of bars: It's likely that the higher the number of bars in a place, the higher the crime rate is likely to be. This is dependent on "nightlife" - there is a higher probability of crime in places which have a lot of nightlife

What we need to show:

After your model building process, you should include a substantial discussion of omitted variables. Identify what you think are the 5-10 most important omitted variables that bias results you care about. For each variable, you should estimate what direction the bias is in. If you can argue whether the bias is large or small, that is even better. State whether you have any variables available that may proxy (even imperfectly) for the omitted variable. Pay particular attention to whether each omitted variable bias is towards zero or away from zero. You will use this information to judge whether the effects you find are likely to be real, or whether they might be entirely an artifact of omitted variable bias.

6. Conclusion