Lab 3 - Stage 3 Final Draft Report

Alex Dauenhauer, Adara Liao, Luis Zorrilla

August 4, 2018

Contents

Introduction

Objective

Our client, Politician Oski, wants to run for office on a platform of improving the quality of living by reducing the crime rate in North Carolina. To do this, our firm, A^3, has been hired to understand the determinants of crime in North Carolina to inform the policy our client will run on. Through regression model building, our research intends to answer, **What are the key explanatory variables for predicting crime rate in North Carolina and how do we act on these variables to create policy reducing crime rate?** Specifically, we will be looking to see what impact can be made on the crime rate by acting on certain characteristics at the county level. We want to know how to predict the crime rate, but we want to be able to predict it using characteristics that can be affected by policy change to support an actionable campaign for Politician Oski.

EDA

We have been given a dataset describing many characteristics of different counties in North Carolina. We first need to conduct clean up or transformations as necessary prior to building our model. We also want to identify 2-3 key explanatory variables that we can use to build the initial model.

Missing Values in Data

There are some missing values in the dataset. It looks like there are 6 N/A's in each variable, appearing in the last 6 rows. We will remove rows with any N/A's so we only have complete observations.

```
data=data[complete.cases(data),]
str(data)
```

```
## 'data.frame':    91 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 ...
```

From the summary above, it appears that the `prbconv` variable is given as a factor. We also see that we have a duplicate data point somewhere in our data frame. We will remove the duplicates and replace the `prbconv` variable with its levels as a numeric vector.

```
data = unique(data)
data = data[order(data$prbconv),]
p = as.numeric(levels(data$prbconv))
data$prbconv = p[!is.na(p)]
```

Outliers in prbarr and prbconv

Looking at the `prbarr` and `prbconv` variables, we see that the maximum value in each is greater than 1. The `prbarr` variable is defined as the ratio of arrests to offenses, and the `prbconv` variable is defined as the ratio of convictions to arrests, so a value of greater than 1 would imply that there were more arrests than offenses and more convictions than arrests. This would indicate either a highly corrupt police force or a bad data point. Even a value equal to 1 would be suspicious. Let's examine these variables, starting with `prbarr`.

```
data[data$prbarr >= 1,]
```

```
##    county year    crmrte  prbarr prbconv prbpris avgsen      polpc
## 51    115   87 0.0055332 1.09091     1.5     0.5   20.7 0.00905433
##      density   taxpc west central urban pctmin80     wcon     wtuc
## 51 0.3858093 28.1931    1       0     0  1.28365 204.2206 503.2351
##        wtrd     wfir     wser   wmfg  wfed   wsta   wloc mix    pctymle
## 51 217.4908 342.4658 245.2061 448.42 442.2 340.39 386.12 0.1 0.07253495
```

So we only have one data point in `prbarr` that is greater than 1. Now we will do the same examination of the `prbconv` variable.

```
nrow(data[data$prbconv >= 1,])
```

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

So we have 10 data points reporting more convictions than arrests including the point where `prbarr` is also greater than 1. This is over 10% of our dataset so we need to make a decision on how to handle these rather than just delete them. If we recoded them as 1 since this should be the maximum allowed value, then we have 10 data points that say every arrest led to a conviction and one data point which says every offense led to a conviction, and this doesn't sound very reasonable. If we delete them, we will be deleting 10% of our other variables as well. We don't have any reason to believe that the data in our other variables for these rows is bad. Therefore, our decision is to recode any values of `prbconv` and `prbarr` that are greater than 1 as `NA`, thus keeping our other variables intact and not unnecessarily removing a large amount of data from them.

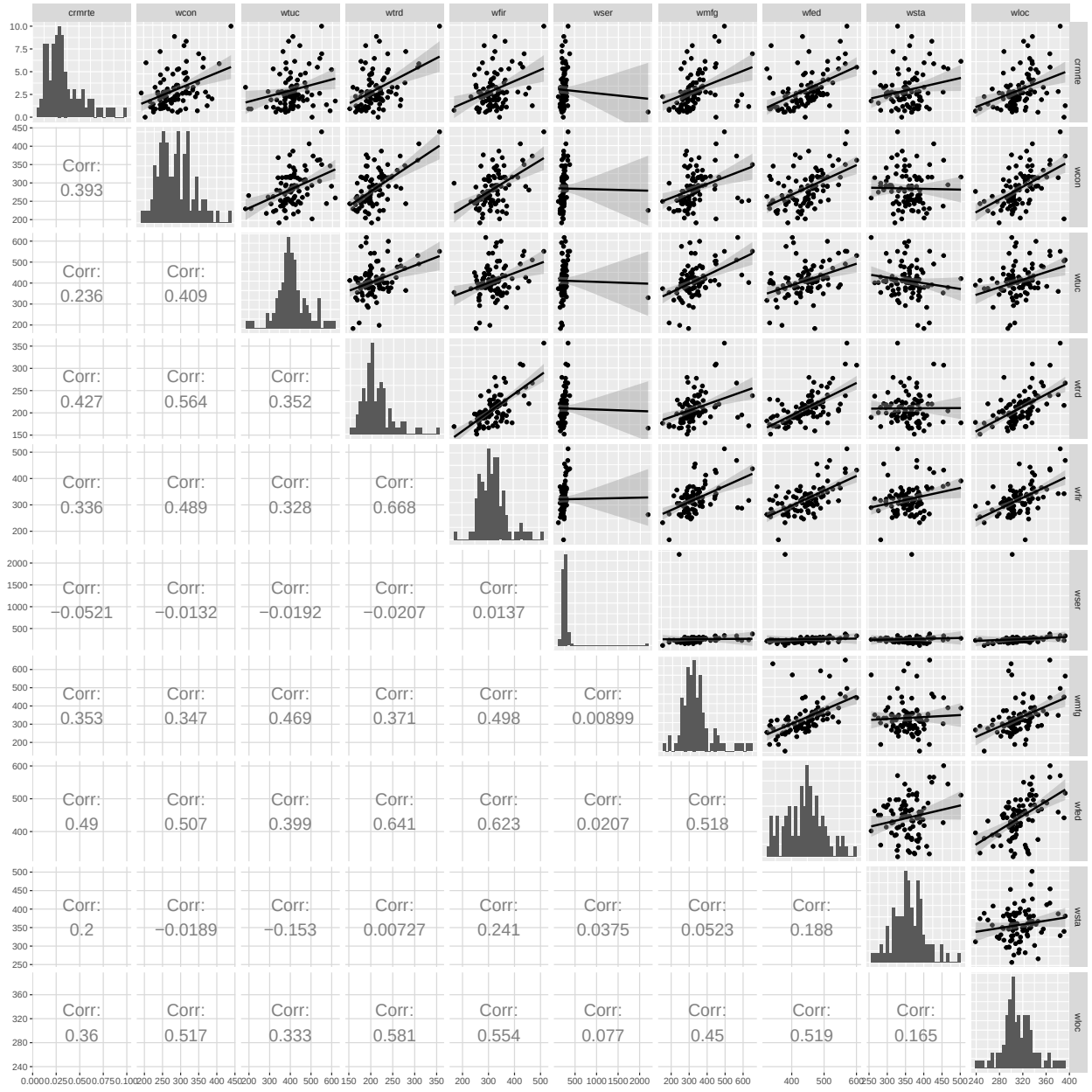
```
data$prbconv[data$prbconv >= 1] = NA
data$prbarr[data$prbarr >= 1] = NA
```

Aggregate Wage Variable

We noticed that there are many weekly wage variables for different industries. As crime is typically linked to low-income areas, let's take a look at how these wage variables are related to the crime rate. We first generate a scatterplot matrix to observe the pattern of correlation among wage variables.

```
cols = c('crmrte', "wcon", "wtuc", "wtrd", "wfir", "wser","wmfg", "wfed",
         "wsta", "wloc")
```

```
ggpairs(data, columns=cols, upper = list(continuous='smooth'),
        lower = list(continuous = wrap(ggally_cor, size=6)),
        diag = list(continuous='barDiag', bins=10))
```



There appears to be a substantial outlier in the `wser` variable. Let's examine that

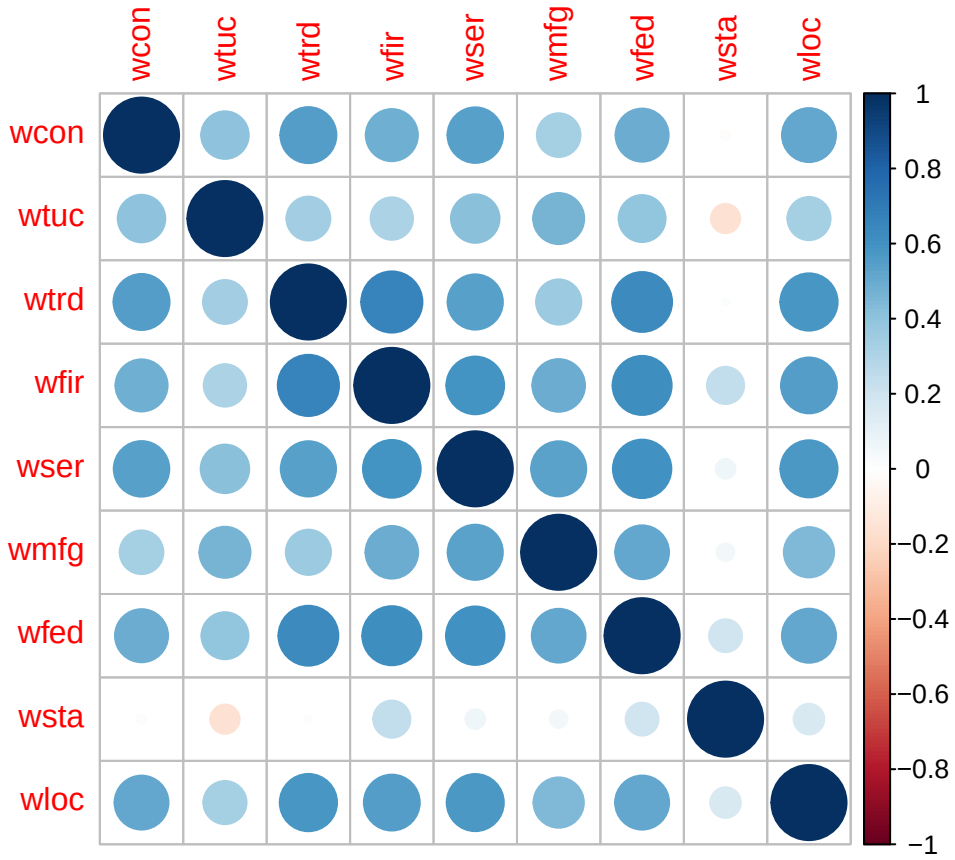
```
data[data$wser == max(data$wser), cols[2:length(cols)]]
```

```
##        wcon    wtuc     wtrd     wfir     wser   wmfg   wfed   wsta   wloc
## 84 226.8245 331.565 167.3726 264.4231 2177.068 247.72 381.33 367.25 300.13
```

It doesn't make any sense that county 84 would have an average weekly wage in the service industry that is an order of magnitude greater than all the other industries. This appears to be an erroneous data point that we will want to remove from our dataset. Rather than simply deleting the point, we may be able to smooth it out by creating an aggregate wage variable across all industries.

We expect the weekly wage variables across sectors to be collinear to some degree, so we would like to examine whether or not we can reasonably aggregate the weekly wage across all industries. To do this we will examine the correlation matrix of weekly wage variables with the outlier point in `wser` temporarily removed:

```
correl = cor(data[!(rownames(data) == 84), cols[2:length(cols)]])
corrplot(correl, method='circle')
```



So with exception to the state employees, weekly wage is fairly collinear across all industries. It makes sense to aggregate this into a single variable and look at the median wage across industries. This will remove any effect from the outlier point and should not be impacted strongly by the lack of correlation in state employees. To ensure that the correlation isn't being affected negatively by including the state employee weekly wage, we can create a second median variable which leaves it out to see what the effect is.

Create median variable `wmedian` for weekly wage

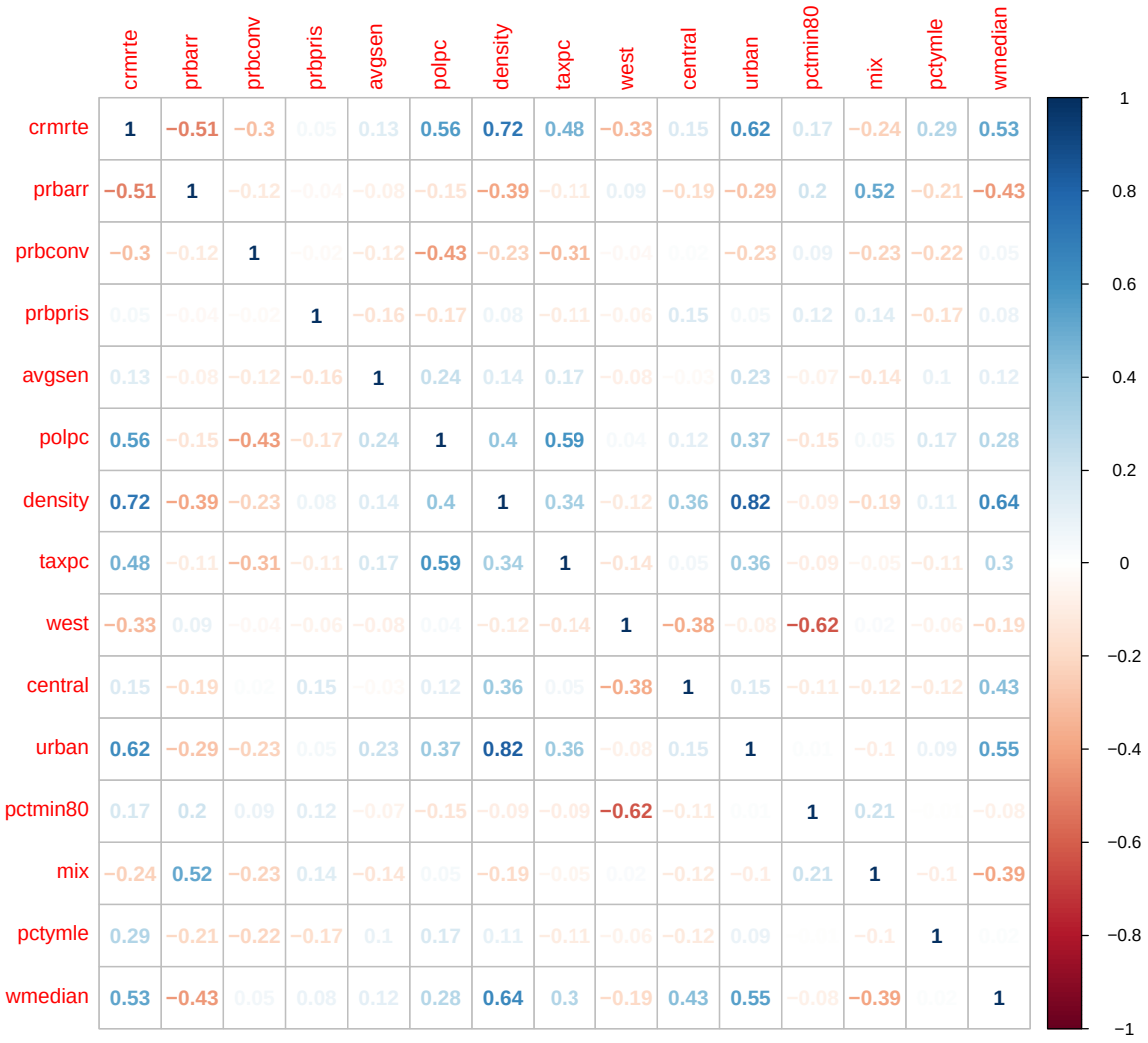
```
data$wmedian = apply(data[cols[2:length(cols)]], MARGIN=1, median)
data$wmedian_noState = apply(data[cols[c(3:9,length(cols))]], MARGIN=1, median)
cor(data[,c('crmrte','wmedian','wmedian_noState')])
```

```
##                    crmrte   wmedian wmedian_noState
## crmrte          1.0000000 0.5109938       0.5001260
## wmedian         0.5109938 1.0000000       0.9614964
## wmedian_noState 0.5001260 0.9614964       1.0000000
```

So the correlation to `crmrte` is nearly identical with or without the state employees so we will leave it in (since we have no other reasons to remove it). Now that we have aggregated our wage variables, let's take a look at how our available predictors correlate to our dependent variable crime rate.

```
cols = colnames(data[c(1:14,24:(length(colnames(data))-1))])
data = data[,cols]
```

```
correl = cor(na.omit(data[3:ncol(data)]))
corrplot(correl, method='number')
```



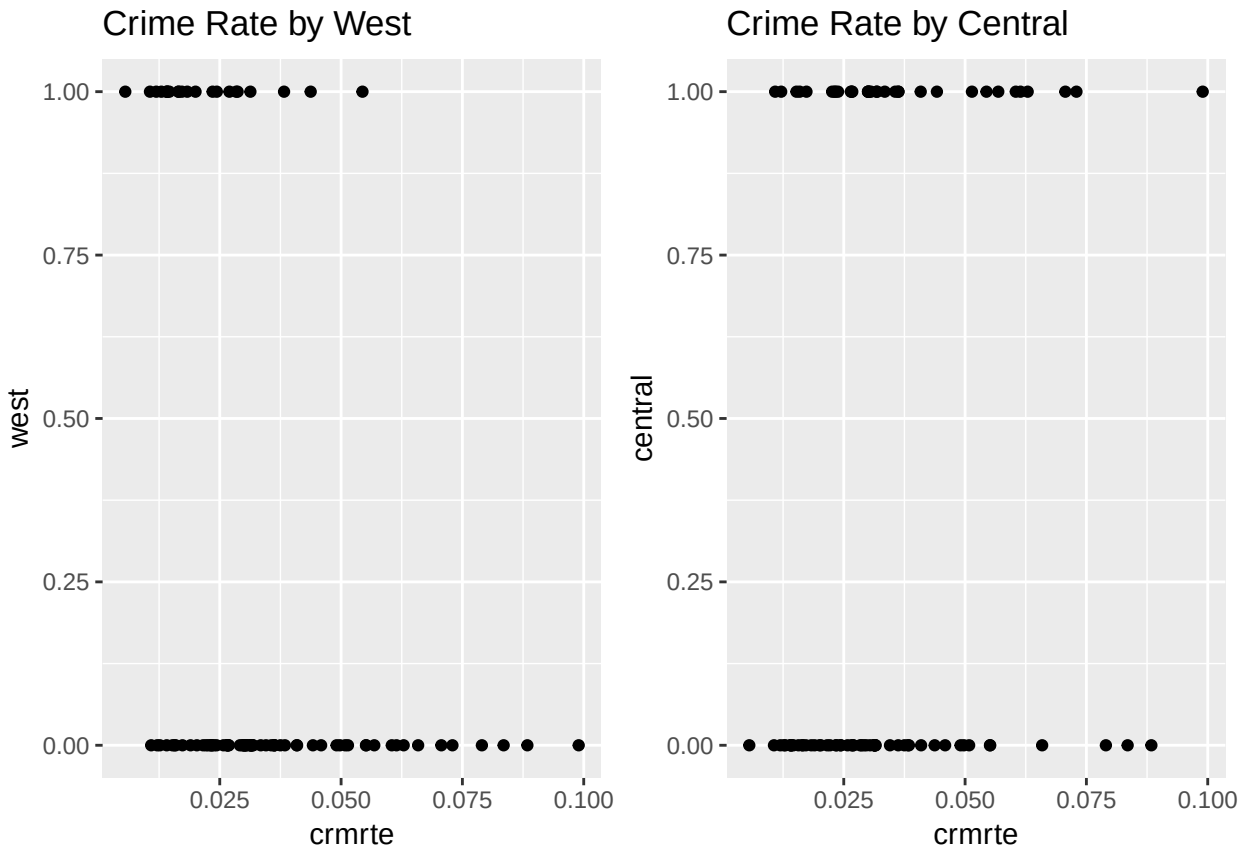
From this we can see that the strongest correlation to crime rate are in `prbarr`, `prbconv`, `polpc`, `density`, `taxpc`, `west`, `urban`, `pctymle` and `wmedian`.

Variables Dropped from Analysis

Geographic variables `west` and `central`

There are 22 entries from west and 34 entries from central.

```
plot_west = ggplot(data, aes(x=crmrte, y=west)) + geom_point() + labs(title="Crime Rate by West")
plot_central = ggplot(data, aes(x=crmrte, y=central)) + geom_point() + labs(title="Crime Rate by Central
grid.arrange(plot_west, plot_central, ncol=2)
```




```
summary(lm(crmrte ~ west + central, data=data))
```

```
##
## Call:
## lm(formula = crmrte ~ west + central, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.027064 -0.011071 -0.005216  0.006569  0.061032
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.036511   0.002989  12.216  < 2e-16 ***
## west        -0.014476   0.004772  -3.034  0.00318 **
## central      0.001423   0.004230   0.337  0.73728
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.01791 on 87 degrees of freedom
## Multiple R-squared:  0.1206, Adjusted R-squared:  0.1003
## F-statistic: 5.963 on 2 and 87 DF,  p-value: 0.003743
```

It appears that locations in west have lower mean crime rate than those not in west. We considered `west` as an indicator variable for our base model, but since this is not actionable for our candidate, we are dropping `west` from consideration. One cannot advocate for lower crime rate by advising people to simply "move west".

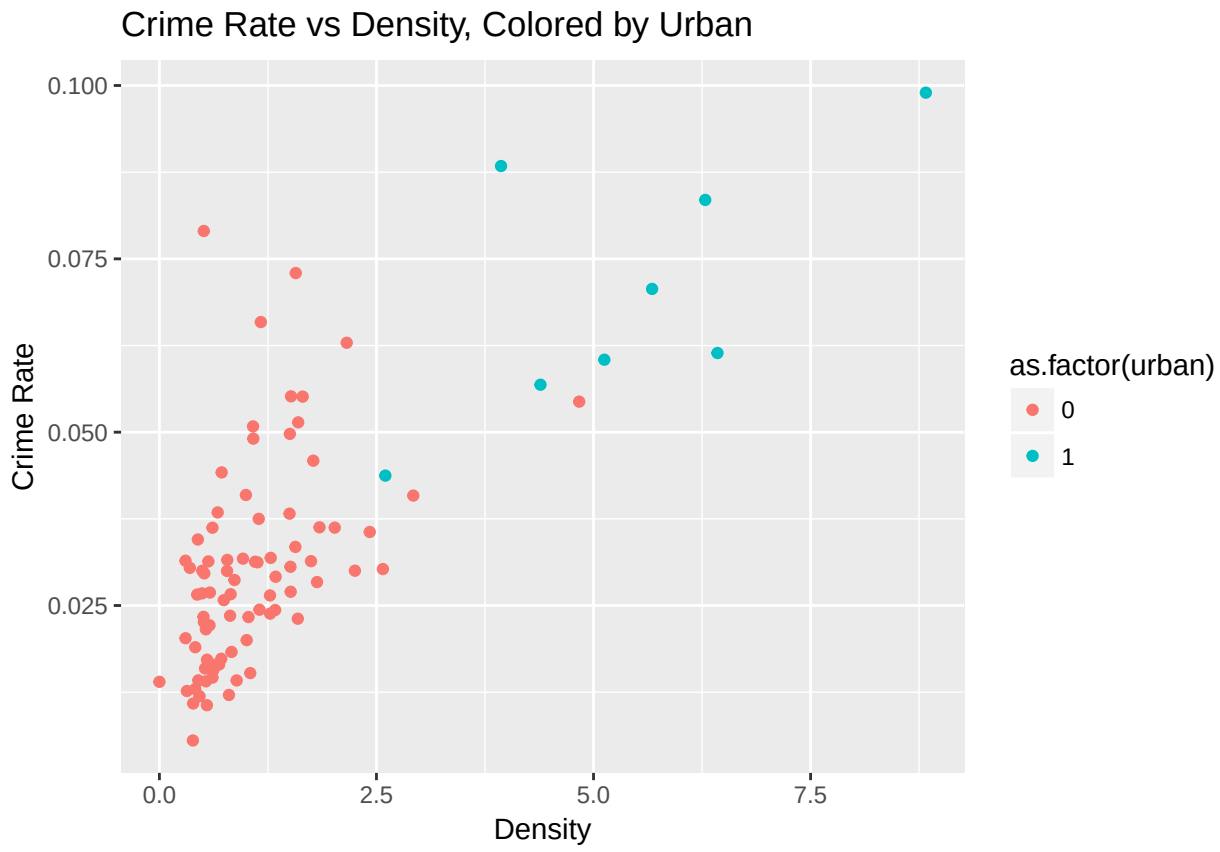
Additionally, `west` is not a desirable indicator variable as it does not have a large impact on `crmrte`. Although statistically significant, `west` has low coefficient, low R^2 and lacks practical interpretation. ("How far west do you have to go in order to reduce crime rate by 0.014 units?") Our client can instead refer to cities in the west as being a desirable model.

Looking at `central`, this is a factor that does not change crime rate prediction by much. `west` is stronger explanatory variable for crime rate than `central`. This can also be seen in stronger correlation of `west` and `crmrte`. Thus, for reasons that are same as `west`, we will not consider `central` as explanatory variable.

`urban` is proxy for `density`

The variable `urban` is essentially a proxy for `density`. From below chart, the density is clearly partitioned along `urban` status, so we will leave `urban` out of analysis. Since the variable `urban` is essentially a proxy for `density`. We will leave the `urban` indicator our of our analysis going forward.

```
ggplot(data, aes(x=density, y=crmrte)) + geom_point(aes(color=as.factor(urban))) + labs(title="Crime Ra
```



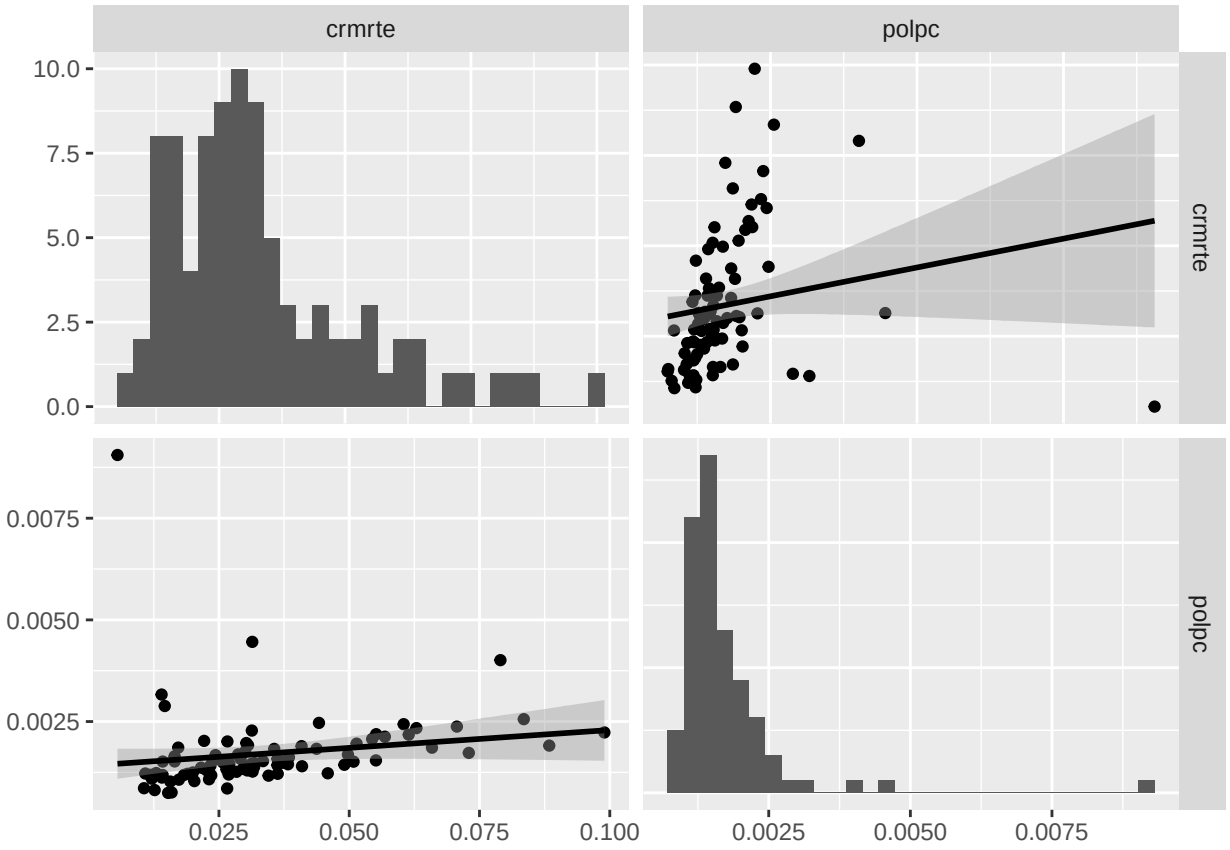
While the geographic variables are helpful to have a better understanding of the spatial distribution of crime, we consider that policy design and implementation around geographic indicator variables is focused

on long-term planning and is not a characteristic that can be directly influenced by politician Oski in his campaign. For these reasons we are not considering the geographic indicator variables.

Police per capita as a response variable rather than explanatory variable

Additionally, we see there is a somewhat strong positive correlation between police per capita and crime rate. This is the opposite effect that we expected. Let's observe a scatterplot matrix for these two variables

```
cols = c("crmrte", "polpc")
ggpairs(data, columns=cols, lower = list(continuous = 'smooth'),
        upper = list(continuous = 'smooth'),
        diag=list(continuous='barDiag', bins=20))
```



So our plot appears much more linear when the crime rate is predicting police per capita rather than vice versa. Let's observe how these two variables predict each other.

```
model_polpc = lm(crmrte ~ polpc, data=data)
model_crmrte = lm(polpc ~ crmrte, data=data)
c1 = coeftest(model_polpc, vcov = vcovHC)
c2 = coeftest(model_crmrte, vcov = vcovHC)
stargazer(model_polpc, model_crmrte, type='latex', report='vcs', header=FALSE,
          keep.stat = c('rsq','adj.rsq'), coef=list(c1,c2),
          omit.table.layout = 'n', table.placement = 'ht')
```

So we see from our coefficient test that `polpc` is not a significant predictor of `crmrte`, and `crmrte` is not a significant predictor of `polpc`. This relationship should be interpreted by thinking that if the crime rate is increasing, we should increase the police force. Otherwise, the logic would be that increasing the police force

Table 1:

	Dependent variable:	
	crmrte	polpc
	(1)	(2)
polpc	3.188	
	(2.003)	
crmrte		0.009
		(0.006)
Constant	0.028	0.001
	(0.004)	(0.0002)
R^2	0.028	0.028
Adjusted R^2	0.017	0.017

will make the crime rate increase which makes no sense. So we will remove `polpc` as a possible predictor of `crmrte` and consider it to be more of a response variable.

Specification for base model

At this point, we are ready to choose the variables we want to build our model with based on correlation to `crmtrte`

Population density: not selected Most difficult from policy perspective as it is difficult to articulate a platform of reducing density. Therefore, we are not include density in our initial model since it is not easily actionable.

Median weekly wage variable: not selected Median weekly wage is correlated positively with crime rate, which is opposite direction to our expectations. Thus we believe that there are some confounding factors hidden within `wmedian`. In theory, wage increase should reduce crime as more economic stability should lead citizens away from depending on crime for income. So we are not going to consider this variable in initial model since the result does not match the theory and we cannot yet explain why.

Variables related to certainty of punishment (`prbarr` and `prbconv`): selected `prbarr` and `prbconv` have a negative correlation with crime rate which is the direction we want

tax per capita: selected Tax revenue per capita is positively correlated with crime rate, which is certainly a platform that a politician can run on!

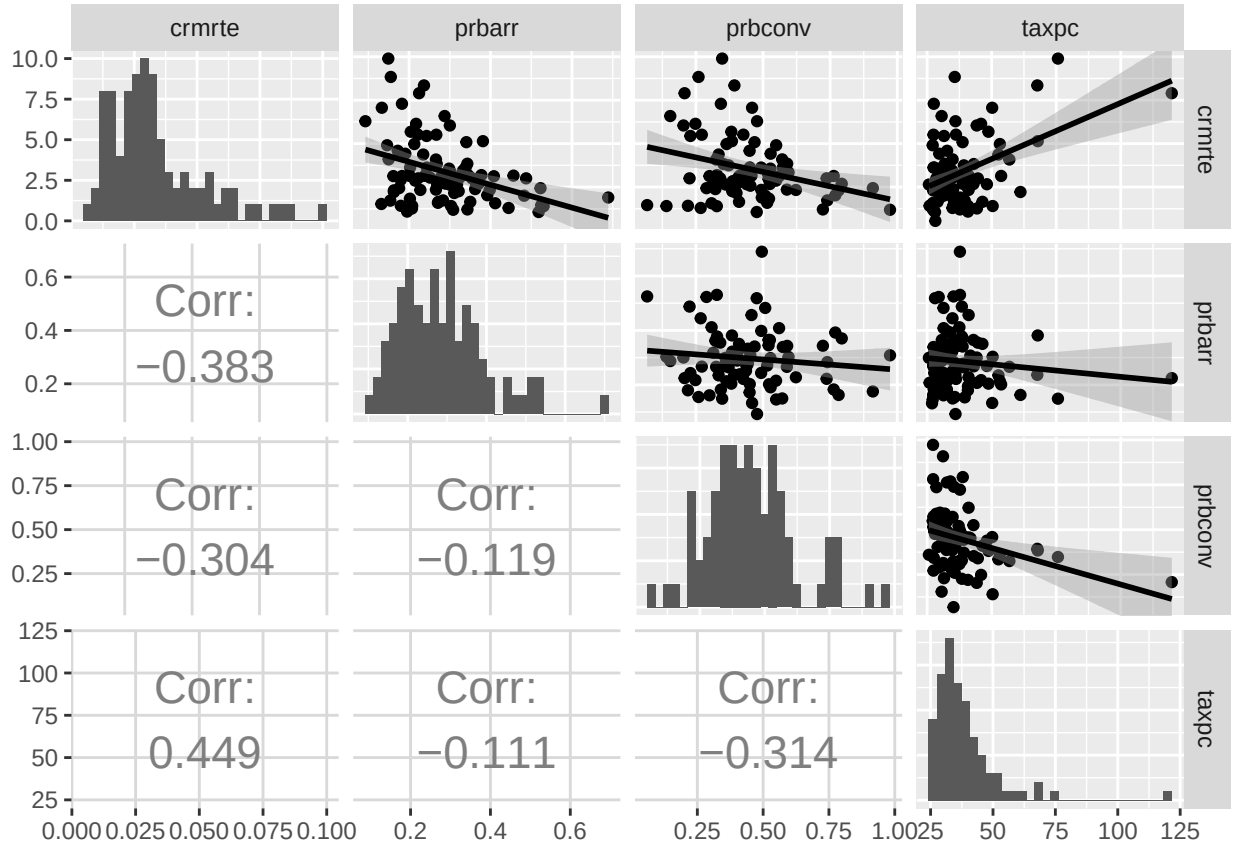
We will build our initial model with the explanatory variables of `prbarr`, `prbconv`, and `taxpc` with the goal of describing how we can lower the crime rate by increasing the "certainty of punishment" and decreasing taxes.

Univariate Analysis of Key Explanatory Variables

Let's now take a look at our the covariates we want to use in our model to see if data transformations may be required

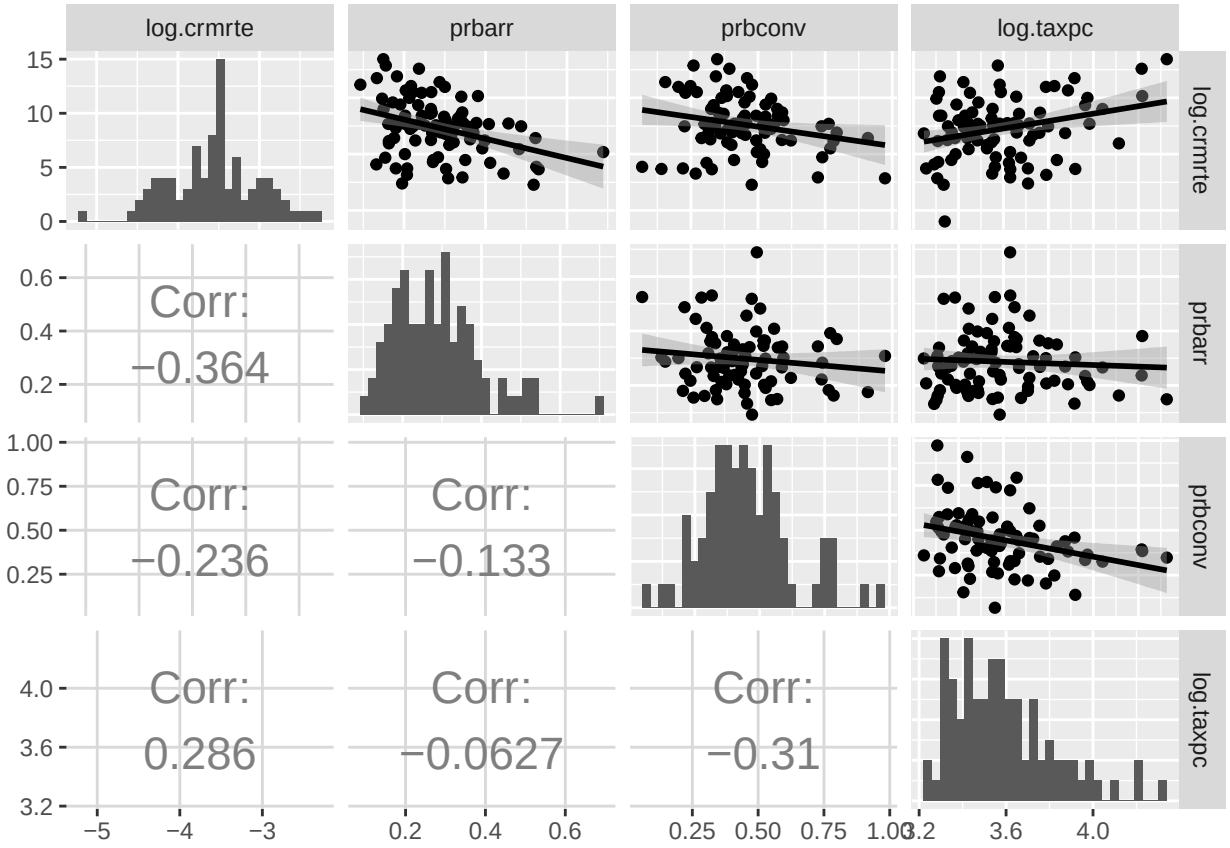
```
cols = c('crmrte', 'prbarr','prbconv','taxpc')
ggpairs(data, columns=cols,
        lower = list(continuous = wrap(ggally_cor, size=6)),
```

```
        upper = list(continuous = 'smooth'),
        diag = list(continuous='barDiag'))
```



We notice a few things here. First we notice that there is an outlier in `taxpc` that is causing a substantial positive skew in this variable. Second, we see that we don't have any perfect collinearity in our covariates (which we also observed in our correlation matrix) so we know going forward that we have not violated CLM 3. We also see that we have a strong positive skew in `crmrte` as well, though this appears to be more representative of the population and not caused by a single outlier. Through pure observation, we can see that the outlier point in `taxpc` has a strong influence on the perceived relationship between `crmrte` and `taxpc`. We want our key explanatory variables to represent the population as closely as possible, so it makes sense to remove this point and observe the relationship to `crmrte` without it. Even without this point, the skew in `crmrte` and `taxpc` may introduce heteroskedasticity issues in our model, so we can take the log of both and see if this reduces the skew.

```
data = data[data$taxpc != max(data$taxpc), ]
data$log.crmrte = log(data$crmrte)
data$log.taxpc = log(data$taxpc)
cols = c("log.crmrte",'prbarr','prbconv','log.taxpc')
ggpairs(data, columns=cols, upper = list(continuous="smooth"),
        lower = list(continuous=wrap(ggally_cor, size=6)),
        diag = list(continuous='barDiag', bins=20))
```



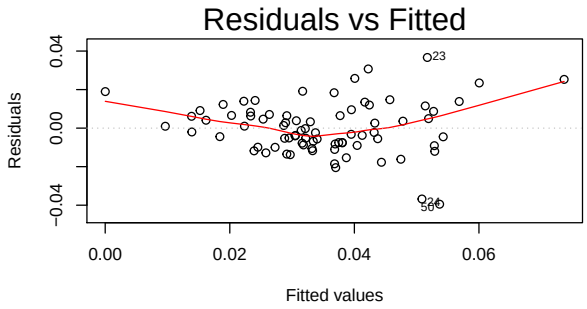
Taking the log of the crime rate appears to eliminate the skew in that variable, but for the tax per capita, it doesn't help that much, and the correlation between our predictors and our response gets worse when we do this. So we may need to do this if we have violated any CLM assumptions, but for now, we will move forward with the variables untransformed.

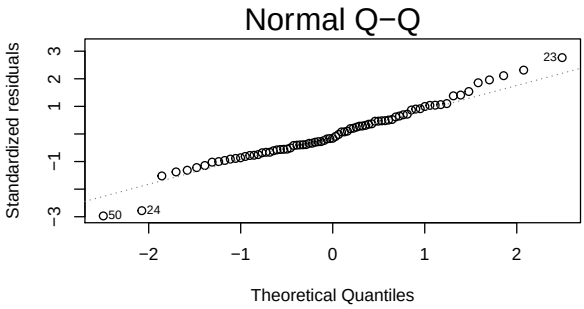
Building the Model

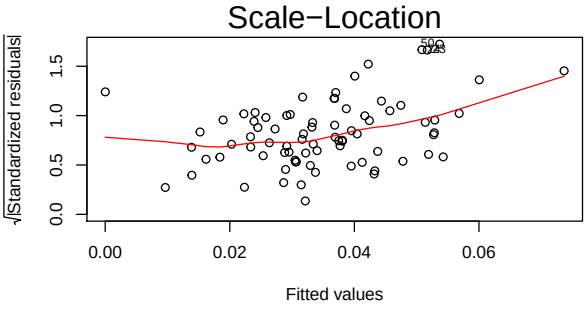
Base Model

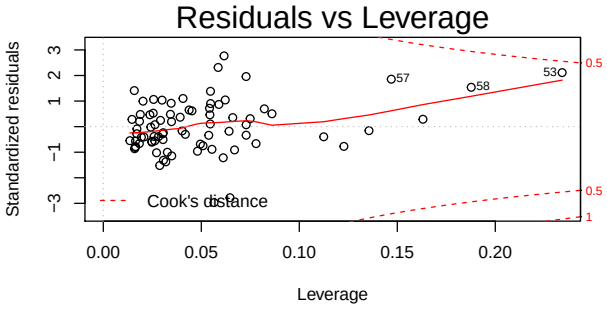
Our first model uses only the key explanatory variables we are focused on. `prbarr`, `prbconv`, and `taxpc`.

```
model1.1 = lm(crmrte ~ prbarr + prbconv + taxpc, data=data)
par(mfrow=c(2,2))
plot(model1.1, which=1, cex.caption = 1.5)
plot(model1.1, which=2, cex.caption = 1.5)
plot(model1.1, which=3, cex.caption = 1.5)
plot(model1.1, which=5, cex.caption = 1.5)
```










```
bptest(model1.1)
shapiro.test(model1.1$residuals)
dwtest(model1.1, alternative='two.sided')
#Conduct Farrah-Glauber test to test for presence of collinearity
imcdiag(data[,c("prbarr", "prbconv", "taxpc")], data[,"crmrte"])
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1.1
## BP = 16.536, df = 3, p-value = 0.0008805
##
##
##  Shapiro-Wilk normality test
##
## data:  model1.1$residuals
## W = 0.9787, p-value = 0.2093
##
##
##  Durbin-Watson test
##
## data:  model1.1
## DW = 2.0247, p-value = 0.9922
## alternative hypothesis: true autocorrelation is not 0
##
##
## Call:
## imcdiag(x = data[, c("prbarr", "prbconv", "taxpc")], y = data[,
##     "crmrte"])
##
```

```
##
## All Individual Multicollinearity Diagnostics Result
##
##            VIF    TOL     Wi     Fi Leamer   CVIF Klein
## prbarr  1.0335 0.9676 1.2732 2.5798 0.9837 1.1406     0
## prbconv 1.1210 0.8921 4.5964 9.3139 0.9445 1.2371     0
## taxpc   1.1081 0.9025 4.1059 8.3199 0.9500 1.2228     0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.4677
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

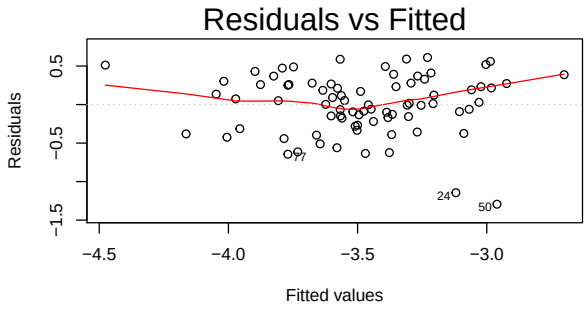
Testing CLM Assumptions

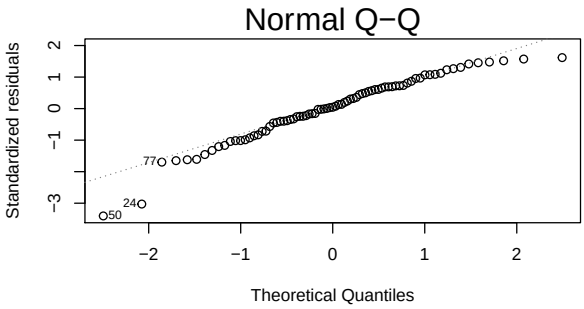
1. Linear in Parameters Assumption is met. We can always draw a line of best fit if we don't care about the size of the errors.
2. Random Sampling The Durbin-Watson test tells us whether or not we see autocorrelation among our explanatory covariates. We do not get a significant result from this test, so with this fact combined with observation of the distributions of the variables above showing no signs of problematic clustering, we determine that we have met CLM assumption 2, Random Sampling.
3. No perfect collinearity From the Farrar-Glauber test, there is no collinearity detected, so MLR.3 assumption is met.
4. Zero conditional mean We can see that endogeneity has occured in the Residuals vs. Fitted Values plot. The expected value of the mean does not appear to remain at a constant zero-value as we move left to right in the plot.
5. Homoskedasticity Looking at the Residuals vs. Fitted Values plot and the Scale-Location plot, we see that the variance of the errors does not appear to be constant for all values. The result of our Breusch-Pagan test confirms that we do have heteroskedasticity in our model so we have violated CLM 5.
6. Normality of errors We can see from the results of our Shapiro-Wilk test and upon observation of the Q-Q plot, that we have good normality of errors so we appear to be maintaining CLM 6.

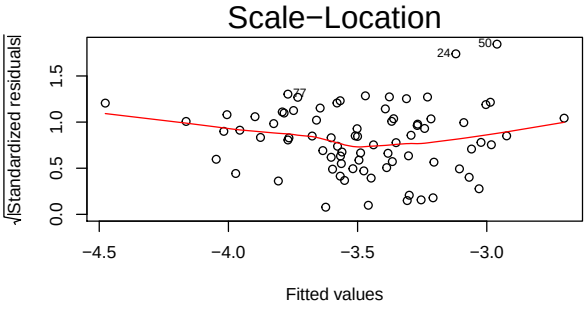
Responding to CLM violations

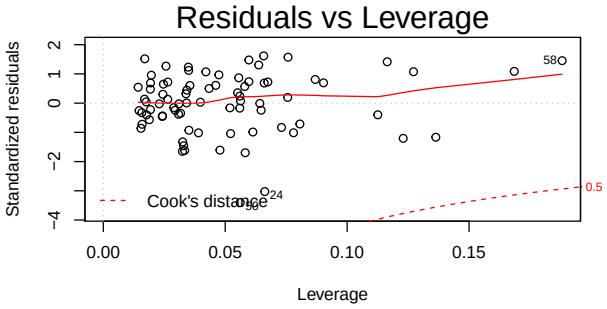
We can attempt to correct the violations of endogeneity and heteroskedasticity by removing some of the skew we saw earlier in `taxpc` and `crmrte` by taking the log of each.

```
model1.2 = lm(log(crmrte) ~ prbarr + prbconv + log(taxpc), data=data)
par(mfrow=c(2,2))
plot(model1.2, which=1, cex.caption = 1.5)
plot(model1.2, which=2, cex.caption = 1.5)
plot(model1.2, which=3, cex.caption = 1.5)
plot(model1.2, which=5, cex.caption = 1.5)
```










```
bptest(model1.2)
shapiro.test(model1.2$residuals)
dwtest(model1.2, alternative = 'two.sided')
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1.2
## BP = 6.0108, df = 3, p-value = 0.1111
##
##
##  Shapiro-Wilk normality test
##
## data:  model1.2$residuals
## W = 0.95538, p-value = 0.007593
##
##
##  Durbin-Watson test
##
## data:  model1.2
## DW = 2.1622, p-value = 0.5332
## alternative hypothesis: true autocorrelation is not 0
```

So it appears we traded one issue for the other. We were able to remove most of the heteroskedasticity we were seeing, but we have now violated our normality of errors assumption. Let's compare our two iterations of this model using heteroskedastic-consistent standard errors of the coefficients.

```
c1 = coeftest(model1.1, vcov=vcovHC)
c2 = coeftest(model1.2, vcov=vcovHC)
stargazer(model1.1, model1.2, type='latex', model.numbers = FALSE, header=F,
          dep.var.labels = c("Crime Rate", "log(Crime Rate)"), omit.stat='ser',
```

```
          title = "Linear Models Predicting Crime Rate in North Carolina",
          covariate.labels = c("Prob. Arrest", "Prob. Conviction", "Tax Rev. per Capita",
                               "log(Tax Rev. per Capita)"),
          coef = list(c1,c2), star.cutoffs=c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(model1.1),1),
                           round(AIC(model1.2),1))), table.placement = 'ht')
```

Table 2: Linear Models Predicting Crime Rate in North Carolina

	Dependent variable:	
	Crime Rate	log(Crime Rate)
Prob. Arrest	-0.087^{***}	-2.519^{***}
	(0.014)	(0.410)
Prob. Conviction	-0.027^{**}	-0.721^{**}
	(0.010)	(0.275)
Tax Rev. per Capita	0.001^{***}	
	(0.0002)	
log(Tax Rev. per Capita)		0.444^{*}
		(0.206)
Constant	0.051^{***}	-3.995^{***}
	(0.010)	(0.812)
AIC	-448	81.8
Observations	79	79
R^2	0.468	0.411
Adjusted R^2	0.446	0.388
F Statistic (df = 3; 75)	21.962^{***}	17.462^{***}
Note:	*p<0.05; **p<0.01; ***p<0.001	

Assessing the Model

So our model got worse when we made this transformation. Our R^2 value decreased, our AIC increased dramatically and our F-statistic decreased. While the interpretation of the log specification might actually be easier to interpret (a percent change in tax revenue results in a percent change in crime rate) we don't want to sacrifice our ability to predict our dependent variable and we can see that taking the log of `taxpc` and `crmrte` will make our ability to predict changes in the crime rate worse. So we want to continue forward with model 1.1, however, we believe we have violated the zero-conditional mean assumption and homoskedasticity. Since we are working with a large dataset ($N = 90 > 30$) we can rely on OLS asymptotics and heteroskedastic-robust standard errors to get around the issue of heteroskedasticity. Since our client cares about building a causal model rather than simply an associative model we need to know the consistent effect of each variable, not just build a line of best fit. Because of this, we cannot yet make the claim that our estimates are consistent since we are seeing evidence of zero-condtional mean violation (endogeneity). We will need to wait until we add more covariates in the next section to see if we improve this and can meet this assumption in our next model specification.

Now we have our preliminary model specification using only our key explanatory variables, but it only explains

45% of the variation in the crime rate. Since `prbarr` and `prbconv` are currently expressed as a percentage, we need to interpret our coefficients a bit differently than they are represented. Our model says that increasing the probability of arrest by 1, will decrease the crime rate by 0.087 crimes per capita, and increasing the probability of conviction by 1 will decrease the crime rate by 0.03 crimes per capita. Increasing `prbarr` and `prbconv` by 1 in this context means increasing them by 100% which doesn't have any practical sense. Instead, we should interpret this model by saying, if we increase the `prbarr` by 1% (0.01) we will decrease the crime rate by 0.00087 crimes per capita. This is roughly a 2.9% decrease in the crime rate when compared to the median crime rate in North Carolina. Likewise, increasing `prbconv` by 1% would result in roughly a 1% decrease in crime rate. For the last variable, assuming that the `taxpc` variable is in the units of $100, we can interpret our model as saying, if we increase the tax revenue per capita by $100, we will increase our crime rate by 0.001 crimes per capita or about 3.3%. This might make sense since increasing the tax revenue means decreasing the income of the residents and typically lower income communities are correlated with a higher crime rate.

Since we have only explained 45% of the variation in the crime rate, at this point we can try to improve the model by including variables from our dataset that we previously omitted.

Improving the Model

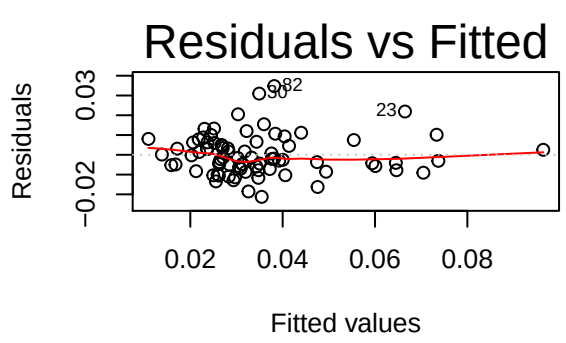
To improve the model and attempt to remove endogeneity or omitted variable bias in our estimator coefficients, we want to start by adding in the covariates that were highly correlated with the crime rate, `density`, `wmedian`, and we will also add the next highest `pctymle`. By adding in these additional variables, our goal is to build a model that is more robust and explains more of the variation in our dependent variable and by doing so, we will also remove omitted variable bias in our key explanatory variables, therefore honing in on the true population parameters. ##Adding `density`, `wmedian`, `pctymle`

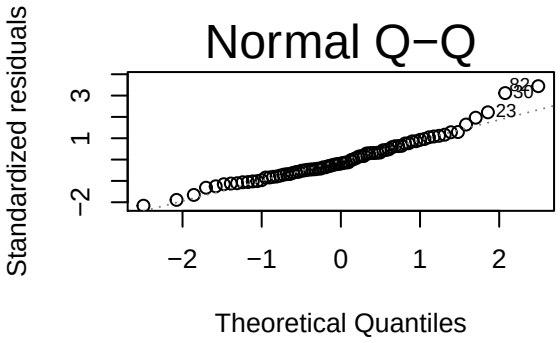
```
model2 = lm(crmrte ~ prbarr + prbconv + taxpc + density + wmedian + pctymle,
            data=data)
bptest(model2)
shapiro.test(model2$residuals)
dwtest(model2)
```

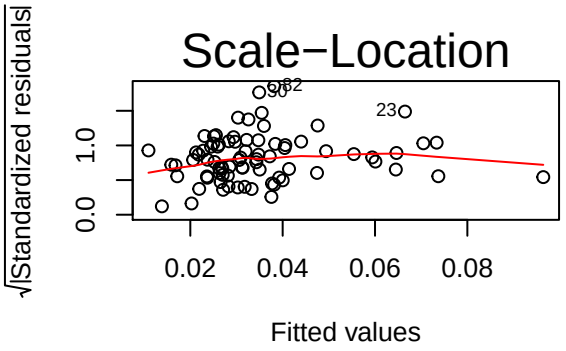
```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 11.602, df = 6, p-value = 0.07147
##
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.95552, p-value = 0.007732
##
##
##  Durbin-Watson test
##
## data:  model2
## DW = 2.2142, p-value = 0.8054
## alternative hypothesis: true autocorrelation is greater than 0
```

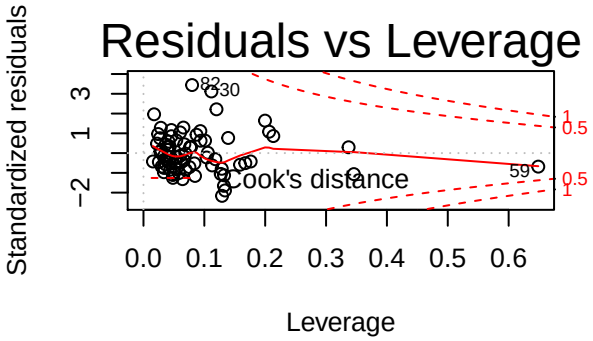
```
par(mfrow=c(2,2))
plot(model2, which=1, cex.caption = 1.5)
plot(model2, which=2, cex.caption = 1.5)
```

```
plot(model2, which=3, cex.caption = 1.5)
plot(model2, which=5, cex.caption = 1.5)
```










```
c2 = coeftest(model2, vcov=vcovHC)
stargazer(model1.1, model2, type='latex', model.numbers = FALSE, header=F,
          dep.var.labels = "Crime Rate", column.labels=c("model1.1", "model2"),
          title = "Linear Models Predicting Crime Rate in North Carolina",
          covariate.labels = c("Prob. Arrest", "Prob. Conviction",
                               "Tax Rev. per Capita", "Pop. Density",
                               "Median Weekly Wage", "Percent Young Male"),
          coef = list(c1,c2), star.cutoffs=c(0.05,0.01,0.001), omit.stat='ser',
          add.lines=list(c("AIC", round(AIC(model1.1),1),
                           round(AIC(model2),1))), table.placement = 'ht')
```

So with the result of our Durbin-Watson test showing that we have not violated our random sampling assumption with the addition of our new covariates, we see that our first three CLM assumptions are still intact. We next notice from our Residuals vs. Fitted Values plot, that our zero-conditional mean assumption appears to be restored. We have a nice flat expectation line at approximately zero. We also notice that we now have violated homoskedasticity at the $\alpha = 0.1$ significance level and normality of errors at the $\alpha = 0.05$ significance level. Since we have a large sample size, under OLS asymptotics we can use heteroskedastic-consistent standard errors on our coefficients and ignore the violation of homoskedasticity in our sample. Using the central limit theorem, we can also ignore the normality of errors violation, since we know these will become normal as our sample size goes to infinity. So this version of our model meets the CLM assumptions we care about under OLS asymptotics and improves the model. Our R^2 value went up, which is to be expected when adding new covariates, but our AIC value went down and our F-statistic went up which shows that we have improved the model with these additions.

Table 3: Linear Models Predicting Crime Rate in North Carolina

	Dependent variable:	
	Crime Rate	
	model1.1	model2
Prob. Arrest	-0.087^{***}	-0.039^{**}
	(0.014)	(0.013)
Prob. Conviction	-0.027^{**}	-0.012
	(0.010)	(0.008)
Tax Rev. per Capita	0.001^{***}	0.0001
	(0.0002)	(0.0002)
Pop. Density		0.006^{***}
		(0.001)
Median Weekly Wage		0.00003
		(0.00004)
Percent Young Male		0.142^{**}
		(0.054)
Constant	0.051^{***}	0.016
	(0.010)	(0.016)
AIC	-448	-486.5
Observations	79	79
R^2	0.468	0.697
Adjusted R^2	0.446	0.672
F Statistic	21.962^{***} (df = 3; 75)	27.604^{***} (df = 6; 72)
Note:		*p<0.05; **p<0.01; ***p<0.001

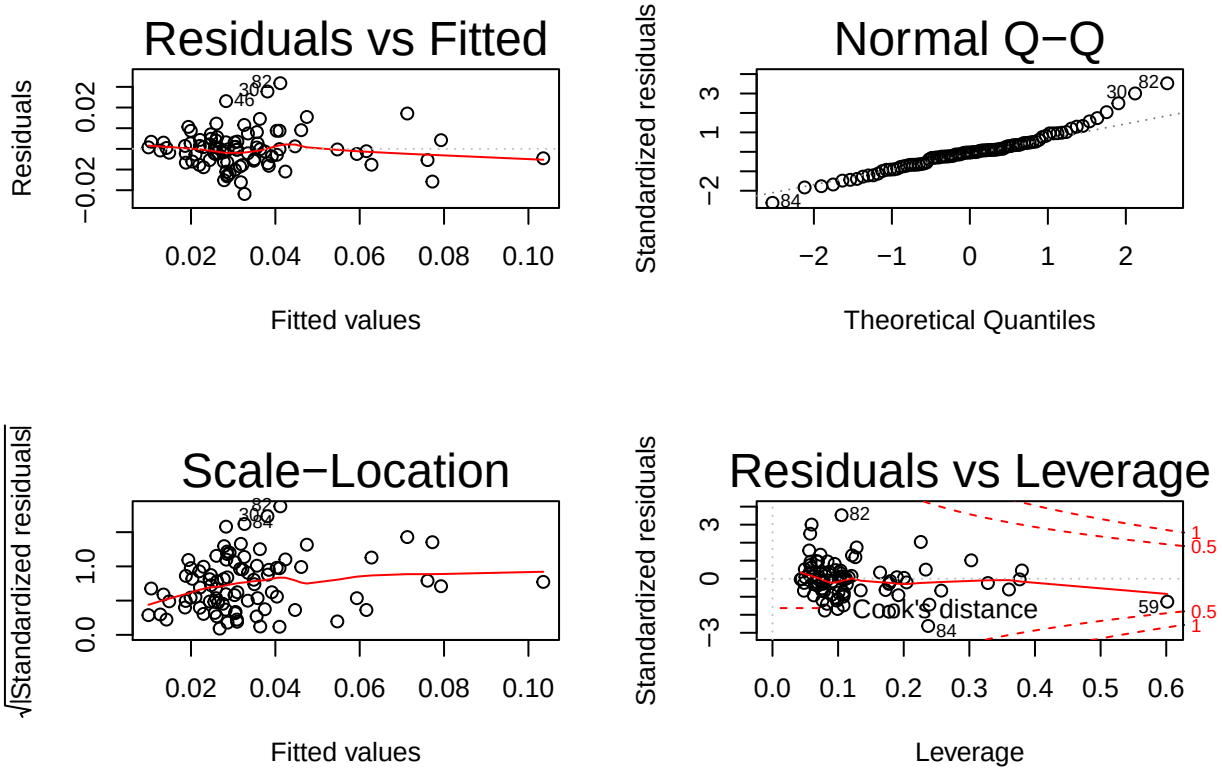
What we notice is that the result of our t-test shows that the weekly median wage does not have a significant causal effect on the crime rate. This is an interesting result since there is so much literature on crime rate being associated with lower income. Looking back at our original correlation matrix, we can see that the weekly median wage is highly correlated with population density. So it is likely that `density` is the true predictor here and the correlation between median weekly wage and crime rate was being confounded by the high collinearity between median weekly wage and density. We also notice that two of our key explanatory variables no longer show to be statistically significant predictors of crime rate and instead our significant predictors are `prbarr`, `density` and `pctymle`. It appears that `prbarr` was being substantially biased by omitting these other variables. The same argument we made for median weekly wage likely also applies to the tax revenue per capita, but we also notice the strong correlation between tax revenue per capita and police per capita. We have already established that police per capita is best viewed as a response to the crime rate. Therefore, logic would imply that the local government is forced to increase taxes to pay for the increase in police force when the crime rate is high. This would lead to the conclusion that `taxpc` is a response to police per capita, meaning it is also a response to the crime rate rather than a predictor. Interpreting why the conviction rate is not a significant predictor is harder to do. We don't see any trends in the correlation matrix that lead us to believe we were confounding our variable with the omission of these other covariates. We may have other omitted variables biasing the probability of conviction, such as recidivism rate, but we will address this more in the omitted variables discussion later on. Finally, we see that the percentage of males

in 1980 has a substantial impact on the crime rate, where a 1% increase in the percentage of young males results in an increase in the crime rate of 0.142 crimes per capita which is roughly 470%. Unfortunately this is not something we can affect with policy changes, but it is good to keep in mind as a significant predictor of the crime rate.

Regression Model Three

At this point we have improved the accuracy and validity of our model, we want to add in the rest of the covariates provided in our dataset to determine whether or not we missed any strong predictors and are introducing omitted variable bias. The alternative would be that we have built a strong model and the covariates that we have omitted from our model thus far, do not really contribute to the prediction of the crime rate. Since our previous model showed that the median weekly wage, tax per capita and probability of conviction, were not statistically significant predictors of the crime rate, we can leave them out of this iteration of the model.

```
model3 = lm(crmrte ~ prbarr + density + pctymle + prbpris + avgsen + west +
              central + urban + pctmin80 + mix, data=data)
par(mfrow=c(2,2))
plot(model3, which=1, cex.caption = 1.5)
plot(model3, which=2, cex.caption = 1.5)
plot(model3, which=3, cex.caption = 1.5)
plot(model3, which=5, cex.caption = 1.5)
```




```
bptest(model3)
shapiro.test(model3$residuals)
```

```
dwtest(model3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 17.412, df = 10, p-value = 0.06572
##
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.95213, p-value = 0.002607
##
##
##  Durbin-Watson test
##
## data:  model3
## DW = 1.8312, p-value = 0.2066
## alternative hypothesis: true autocorrelation is greater than 0
```

```
c3 = coeftest(model3, vcov=vcovHC)
stargazer(model1.1, model2, model3, type='latex', model.numbers = FALSE,
          dep.var.labels = "Crime Rate", header=F, omit.stat='ser',
          column.labels=c("model1.1", "model2", "model3"),
          title = "Linear Models Predicting Crime Rate in North Carolina",
          covariate.labels = c("Prob. Arrest", "Prob. Conviction",
                               "Tax Rev. per Capita", "Pop. Density",
                               "Median Weekly Wage", "Percent Young Male",
                               "Prob. Prison", "Avg. Sentence", "West",
                               "Central", "Urban","Percent Minority 1980",
                               "Offense Mix"),
          coef = list(c1,c2, c3), star.cutoffs=c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(model1.1),1),
                           round(AIC(model2),1), round(AIC(model3),1))))
```

So our first three CLM assumptions are intact as we do not show any statistical significance of autocorrelation among our covariates. We appear to have a fairly straight expectation line hovering around zero in our residuals vs fitted values plot, so it appears that our zero-conditional mean assumption is also met. The result of our Shapiro-Wilk test tells us that we have not met the normality of errors assumption and we see from our Breusch-Pagan test that we have violated our homoskedasticity assumption. This is ok for our model since we are relying on OLS asymptotics and we are using heteroskedastic-consistent standard errors. So our conclusion is that we have met the four CLM assumptions we care about under OLS asymptotics and we can see that the only significant predictors we have in our data set are `prbarr`, `density`, `pctymle`, and `pctmin80`. Our final iteration of our model will include only these variables in an effort to reduce the standard errors of our coefficients.

```
model4 = lm(crmrte ~ prbarr + density + pctymle + pctmin80, data=data)
bptest(model4)
shapiro.test(model4$residuals)
dwtest(model4)
imcdiag(data[,c("prbarr", "density", "pctymle", "pctmin80")], data[,"crmrte"])
```

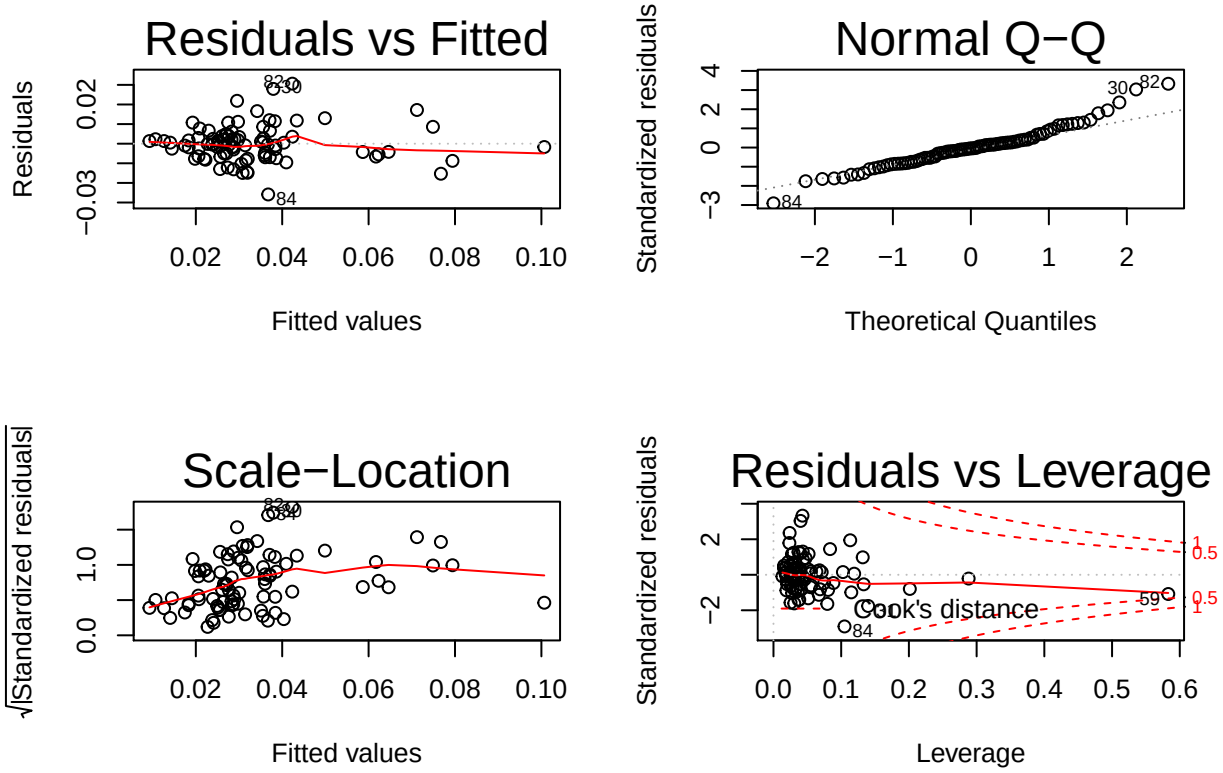
```
##
##  studentized Breusch-Pagan test
```

Table 4: Linear Models Predicting Crime Rate in North Carolina

	Dependent variable:		
	Crime Rate		
	model1.1	model2	model3
Prob. Arrest	-0.087^{***}	-0.039^{**}	-0.026^{*}
	(0.014)	(0.013)	(0.012)
Prob. Conviction	-0.027^{**}	-0.012	
	(0.010)	(0.008)	
Tax Rev. per Capita	0.001^{***}	0.0001	
	(0.0002)	(0.0002)	
Pop. Density		0.006^{***}	0.009^{***}
		(0.001)	(0.001)
Median Weekly Wage		0.00003	
		(0.00004)	
Percent Young Male		0.142^{**}	0.149^{***}
		(0.054)	(0.045)
Prob. Prison			0.008
			(0.013)
Avg. Sentence			-0.0004
			(0.0004)
West			-0.005
			(0.004)
Central			-0.005
			(0.003)
Urban			-0.002
			(0.007)
Percent Minority 1980			0.0002^{*}
			(0.0001)
Offense Mix			0.005
			(0.015)
Constant	0.051^{***}	0.016	0.012
	(0.010)	(0.016)	(0.010)
AIC	-448	-486.5	-557
Observations	79	79	88
R^2	0.468	0.697	0.758
Adjusted R^2	0.446	0.672	0.727
F Statistic	21.962^{***} (df = 3; 75)	27.604^{***} (df = 6; 72)	24.119^{***} (df = 10; 77)
Note:			*p<0.05; **p<0.01; ***p<0.001

```
##
## data:  model4
## BP = 15.189, df = 4, p-value = 0.004326
##
##
##  Shapiro-Wilk normality test
##
## data:  model4$residuals
## W = 0.96215, p-value = 0.01142
##
##
##  Durbin-Watson test
##
## data:  model4
## DW = 1.8117, p-value = 0.1781
## alternative hypothesis: true autocorrelation is greater than 0
##
##
## Call:
## imcdiag(x = data[, c("prbarr", "density", "pctymle", "pctmin80")],
##     y = data[, "crmrte"])
##
##
## All Individual Multicollinearity Diagnostics Result
##
##             VIF    TOL     Wi     Fi Leamer   CVIF Klein
## prbarr   1.1848 0.8440 5.1736 7.8528 0.9187 2.4368     0
## density  1.1292 0.8855 3.6189 5.4930 0.9410 2.3226     0
## pctymle  1.0408 0.9608 1.1418 1.7331 0.9802 2.1407     0
## pctmin80 1.0346 0.9666 0.9682 1.4695 0.9831 2.1279     0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.7461
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
par(mfrow=c(2,2))
plot(model4, which=1, cex.caption = 1.5)
plot(model4, which=2, cex.caption = 1.5)
plot(model4, which=3, cex.caption = 1.5)
plot(model4, which=5, cex.caption = 1.5)
```




```
c4 = coeftest(model4, vcov=vcovHC)
stargazer(model1.1, model2, model3, model4, type='latex', model.numbers = FALSE,
          dep.var.labels = "Crime Rate",  header=FALSE, omit.stat=c('ser','f'),
          column.labels=c("model1.1", "model2", "model3", "model4"),
          title = "Linear Models Predicting Crime Rate in North Carolina",
          covariate.labels = c("Prob. Arrest", "Prob. Conviction",
                               "Tax Rev. per Capita", "Pop. Density",
                               "Median Weekly Wage", "Percent Young Male",
                               "Prob. Prison", "Avg. Sentence", "West",
                               "Central", "Urban","Percent Minority 1980",
                               "Offense Mix"),
          coef = list(c1,c2, c3,c4), star.cutoffs=c(0.05, 0.01, 0.001),
          add.lines=list(c("AIC", round(AIC(model1.1),1), round(AIC(model2),1),
                           round(AIC(model3),1), round(AIC(model4),1))))
```

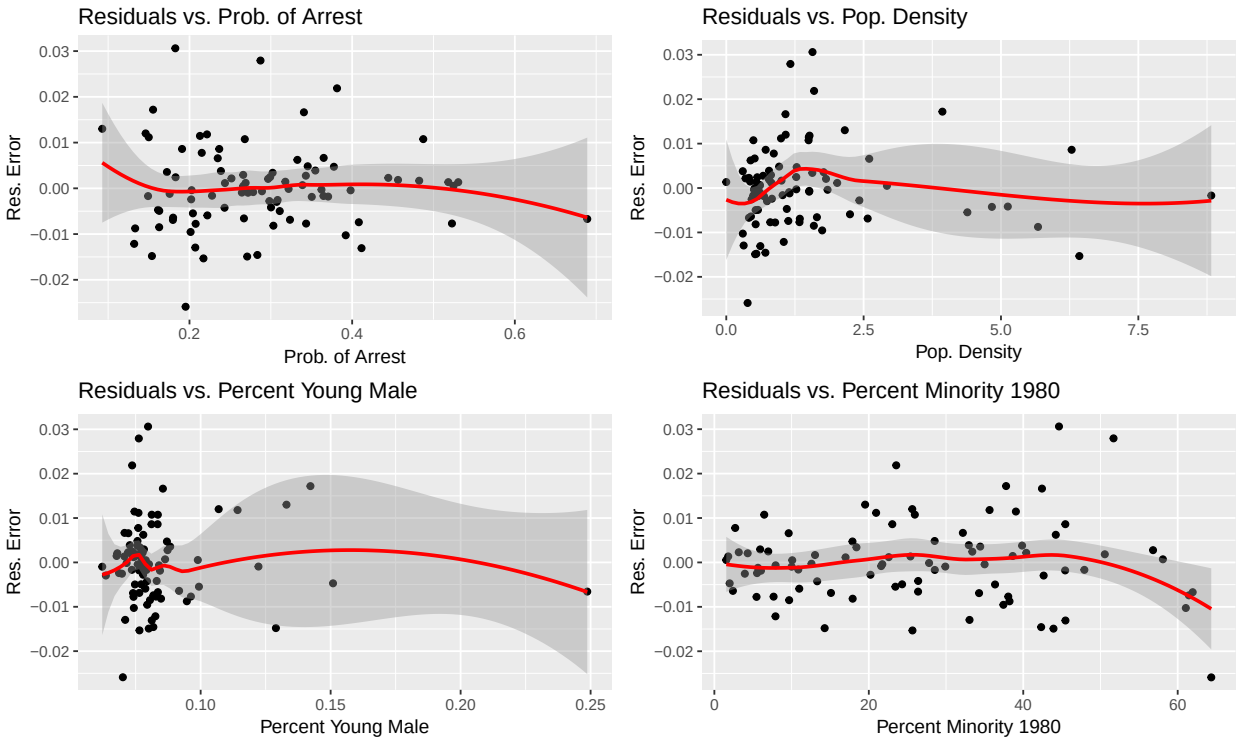
Our final model explains about 74% of the variation in the crime rate in North Carolina. Our AIC is the lowest of all four models which says that we have improved our model while adding covariates from the initial key explanatory variables. We also have reduced the standard errors on all our coefficients to the lowest of all the model specifications. We can see that we still have heteroskedasticity in our final model (both visually from the Scale-Location plot and statistically from the Breusch-Pagan test), but as we are using heteroskedastic-consistent standard errors, this is an acceptable violation of this CLM assumption. Our Q-Q plot combined with the result of our Shapiro-Wilk test shows that we have violated the normality of errors assumption, but we accept this violation under the principals of OLS asymptotics which we can apply to our large dataset. We need background knowledge of how the data was collected to truly assess the random sampling assumption. The intent of this assumption is to know that the sample is representative of the population. Since North Carolina has 100 counties and we have data for 90 of them, we seem to be in good

Table 5: Linear Models Predicting Crime Rate in North Carolina

	Dependent variable:			
	Crime Rate			
	model1.1	model2	model3	model4
Prob. Arrest	-0.087^{***}	-0.039^{**}	-0.026^{*}	-0.026^{*}
	(0.014)	(0.013)	(0.012)	(0.010)
Prob. Conviction	-0.027^{**}	-0.012		
	(0.010)	(0.008)		
Tax Rev. per Capita	0.001^{***}	0.0001		
	(0.0002)	(0.0002)		
Pop. Density		0.006^{***}	0.009^{***}	0.009^{***}
		(0.001)	(0.001)	(0.001)
Median Weekly Wage		0.00003		
		(0.00004)		
Percent Young Male		0.142^{**}	0.149^{***}	0.159^{***}
		(0.054)	(0.045)	(0.043)
Prob. Prison			0.008	
			(0.013)	
Avg. Sentence			-0.0004	
			(0.0004)	
West			-0.005	
			(0.004)	
Central			-0.005	
			(0.003)	
Urban			-0.002	
			(0.007)	
Percent Minority 1980			0.0002^{*}	0.0003^{***}
			(0.0001)	(0.0001)
Offense Mix			0.005	
			(0.015)	
Constant	0.051^{***}	0.016	0.012	0.006
	(0.010)	(0.016)	(0.010)	(0.005)
AIC	-448	-486.5	-557	-564.8
Observations	79	79	88	88
R^2	0.468	0.697	0.758	0.746
Adjusted R^2	0.446	0.672	0.727	0.734
Note:				*p<0.05; **p<0.01; ***p<0.001

shape for assuming that we have a random sampling. We are told that the population variables (`pctymle` and `pctmin80`) were drawn from census data which should indicate data that is representative of the population. If we assume that the `density` variable was also collected from census data, this sample should also be ok. The probability of arrest variable is derived from the FBI's Uniform Crime Reports. There is a reason to suspect that this data is not fully representative of the true population numbers with how much reporting there has been on poor police reporting over the last few years. Without this knowledge however, we can only rely on observation and statistical testing. With the Durbin-Watson test showing an insignificant result, we conclude that this model meets the assumption of Random Sampling. From the Farrah-Glauber test, we see there is no perfect collinearity among our covariates. The visual observation of our Residuals vs. Fitted Values plot shows a flat line at the zero-value and this implies that our zero-conditional mean assumption is also valid for this model. To fully investigate this assumption, we should look at the residuals vs. predictor plots for each predictor.

```
r = model4$residuals
par(mfrow=c(2,1))
cols = c('prbarr','density','pctymle','pctmin80')
d = data.frame(na.omit(data[,cols]), r)
g1 = ggplot(d, aes(prbarr, r)) + geom_point() +
  geom_smooth(stat="smooth", col='red') +
  labs(title='Residuals vs. Prob. of Arrest',
       x='Prob. of Arrest', y='Res. Error')
g2 = ggplot(d, aes(density, r)) + geom_point() +
  geom_smooth(stat="smooth", col='red') +
  labs(title='Residuals vs. Pop. Density',
       x='Pop. Density', y='Res. Error')
g3 = ggplot(d, aes(pctymle, r)) + geom_point() +
  geom_smooth(stat="smooth", col='red') +
  labs(title='Residuals vs. Percent Young Male',
       x='Percent Young Male', y='Res. Error')
g4 = ggplot(d, aes(pctmin80, r)) + geom_point() +
  geom_smooth(stat="smooth", col='red') +
  labs(title='Residuals vs. Percent Minority 1980',
       x='Percent Minority 1980', y='Res. Error')
grid.arrange(g1, g2, g3, g4, nrow=2, ncol=2)
```



There isn't any strong evidence that we have violated the zero-conditional mean assumption. The deviations from zero in the Prob. Arrest, Percent Minority, and Percent Young Male plots occur where we have very few points or only a single data point and therefore the deviation from zero shouldn't be interpreted as a dependence on the predictor variable. So we believe we have met this assumption. We also see from our Residuals vs. Leverage plot that we don't have any points in our model with high influence, so we aren't skewing our results based on single outlier points. Therefore, this model acceptably passes all 4 CLM assumptions that we care about under OLS asymptotics. This tells us that our estimators are BLUE and we can now interpret our results.

Our coefficient for Probability of Arrest is -0.026. As stated above, the probability of arrest variable is defined as the ratio of arrests to offenses. Therefore we interpret the coefficient as saying that if we were to increase the probability of arrest by 1% (0.01), we would decrease the crime rate by 0.00026 or roughly 1% when compared to the median crime rate. The density variable is defined as the number of people per square mile and the coefficient of the density variable is 0.009. This can be interpreted as meaning that if we were to increase the population density of 1 person per square mile, we would increase the crime rate by 0.009 or roughly 30%. This is a substantial impact on the crime rate. This implies that our candidate may want to come up with policies to reduce the population density, such as better low-income housing or better housing development plans. The percent young male variable tells the percent of the population that is male and between the ages of 15 and 24. The interpretation of its coefficient is that if you increase the percent of young males by 1%, you will increase the crime rate by 0.159 or roughly 530%. This is not something our client can take action on, as we can't really implement policy to reduce the number of young males in the population under current U.S. law. Therefore this variable is included merely to inform the model and remove any omitted variable bias from our actionable predictors. The next variable is the percent of the population that is a minority in 1980. This coefficient can be interpreted as saying that if we had a higher percentage of minorities in 1980 by 1%, we will have a higher crime rate in 1987 by 0.0003 or roughly 1%. In the same fashion as with percent young male, this variable informs the model but is not directly actionable. The last coefficient we need to interpret is our intercept. This comes out with a value of 0.006 crimes per person, which means that with all our predictors held to zero, we still would have a crime rate of 0.006 crimes per person. This makes it clear that we are omitting some key predictor variables and that we have not found the full causal model to explain the crime rate. Now we will discuss some theoretical omitted variables that would likely help improve the explanatory power of our model.

Ommitted Variables

From the list of variables that affect crime according to the FBI, https://ucr.fbi.gov/hate-crime/2011/resources/variables-affecting-crime, we can identify a few variables that would likely influence our crime rate that we have omitted from our model.

Climate: There have been studies that show that temperature is positively correlated to the crime rate. Assuming this relationship holds in North Carolina, since there was no climate data included in our data set, we have introduced a positive bias into our coefficients. The true influence of our model variables is likely smaller than what we are predicting. Therefore, our candidate would need to compensate for that if using our model to enact policy decisions.

Gun Ownership: Gun ownership probably wouldn't affect petty crime, and we don't really have data telling us the percentage of crime that is violent or petty here. So, assuming there is a reasonable percentage of the crime in North Carolina that is violent, gun ownership would likely have a positive correlation on with the crime rate. So we are again introducing a positive bias into the variables we have described in our model and we are again overestimating the influence that our variables will have on the crime rate.

Recidivism: The effect of omitting the recidivism rate would mainly be on the `prbconv` variable in our model. Omitting this data may explain why we don't see the probability of conviction as a significant predictor. A high recidivism rate would certainly increase the crime rate, but it would also likely have a negative bias on the probability of conviction. This is because as we see the recidivism rate go up, this is a strong indicator that serving a sentence is not deterring criminals from committing crime again.

Job Stability: Our initial conclusion without any data pushes us to think that as job stability decreases, crime rate will increase. A good example of this logic is the Oil & Gas industry where 90% of the jobs depend on a highly volatile variable which is the price of the commodity. In the specific regions where the Oil & Gas industry is the primary source of income as in West Texas, crime rate increases as soon as crisis hits and the working population lose their jobs. The lack of job stability also has an impact on long term planning (career, housing, etc) which makes not only the job but the individuals stability suffer that in turn as mentioned above, we believe, job stability has a negative correlation impact in crime rate. This conclusion leads us to believe that our coefficients are smaller than they would be if were to introduce data for this variable into our model.

Ex-Military: While we didn't come up with an agreement on effect size of this omitted variable, we identified Ex-Military Service as one of the omitted variables that could possibly impact crime rate. Our initial conclusion was that based on the well-publicized problems faced by veterans, like mental health issues and challenges transitioning to a regular life, we concluded that the offenses are more severe and the execution of crime is more drastic. Our research group, A^3 believes that in the ideal scenario of having this data the stigma could be denied or confirmed and if proven, better policy strategies could be implemented to aid veterans and reduce crime.

Unemployment Rate: Unemployment rate is not found in our data set and we are unable to find proxy measures for unemployment rate. Unemployment impacts on crime rate due to poverty-motivated crimes. The wage data could have correlation with unemployment (higher unemployment can have downward pressure on wage growth), but we need temporal data which is missing from data set. As we expect wages to be negatively correlated to unemployment, and wages have positive coefficient in the regression model, then OLS coefficient for wages will be scaled towards zero (less positive), losing statistical significance.

Conclusions

Through our regression model analysis, we found four variables that explain 74% of the variation in the crime rate. Our initial assessment of the key explanatory variables turned out to be misguided since our initial focus was to answer a question that already implied predictors based on our own assumptions and biased

perceptions about crime. This was taking our report more to a confirmation/denial exercise rather than a research to identify key variables from diverse data to inform about policy changes to reduce crime.

The six main conclusions of our firm A^3, are presented in this section and additional conclusive material is presented in the following section Comments/Thoughts for curious minds.

1) **Police Effiency** Police force is an asset that needs funding and our conclusion is that since police presence is a consequence of the crime rate, in order to mitigate the need for increased police presence, police has to become more efficient, through policy implementation to increase the probability of conviction and to increase the probability of arrest to deter crime. By doing this, politician Oski would not only reduce the crime rate but also have additional resources for his political campaign.

2) **Density awareness** As found in the research, density is a high predictor of crime and while in the short-term, we do not recommend to run on a density reducing policy, this predictor presents an opportunity for long term planning to improve quality of living by implementing policies to redistribute density to mitigate the effects of highly populated areas and reduce the crime rate not by punishing but by territory management. Therefore, in highly dense areas, the long-term goal would be to encourage movement to less-populated areas by, for example, creating better opportunities to low-income housing access while in short term, promoting more arrests and higher conviction rates.

3) **Wages is not a significant predictor of crime rate** Despite common conception that higher wages depresses crime, the improvement of our model revealed that wages is not significant predictor of crime rate. Probably because wages is highly correlated with density, so omitting density was confounding coefficient of wages variable.

4) **Educate young males** This is a predictor over which we have no control in the present. While in the near future gender selection through policy could be an option, in the current scenario the best our firm can to is to inform about the strong influence that the number of young males have in the crime rate to let politician Oski to decide on policies to educate, inform and motivate crime reduction in young males in North Carolina.

5) **Minority awareness** In order to work on a broad campaign that has both long and short term goals, our firm found that Minorities play an important role in the crime rate. Our actionable conclusions are focused on police efficiency, which could help reduce the impact that this predictor has. Our expectation is that the reason this variable is a successful predictor of the crime rate is that minority communities are overpoliced and overrepresented in the justice system. Better policies for police efficiency could help reduce this result.

6) **Police is a response to crime.** Police force is an asset that needs funding and our conclusion is that since police presence is a consequence of the crime rate, in order to mitigate the need for increased police presence, police has to become more efficient, through policy implementation to increase the probability of conviction and to increase the probability of arrest to deter crime. By doing this, politician Oski would not only reduce the crime rate but also have additional resources for his political campaign.

Comments/Thoughts

This iteration of our report does not match our first draft as we learned additional tools for model specification and testing CLM assumptions after the first stage. Additionally, our methodology was not accurate. We started by choosing our explanatory variables before performing our EDA because we thought we needed to choose them in our research question. This time, we let EDA guide us in identifying key variables. Our first draft led us to a model that very poorly predicted the crime rate adn used variables that we believe were better described as predictors. We decided as a team that we could produce a better result if we started our model building process over. Some of the content is the same, but our model specifications are completely different between the two drafts.