w203_lab3

April 18, 2020

1 Lab 2 - W203 - Statistics for Data Science

Submission by Jude Kavalam, Harshit Patel, Abhi Sharma

1.1 1. Introduction

1.1.1 Dataset

The database contains only a subset of information drawn from a larger multi-year database from a study performed by Cornwell and Trumball researchers from the University of Georgia and the West Virginia University (CT1994). From the database the subset of twenty five variables encompass roughly three types of variables, penal system, economic and population related variables. Each observation is from a different county in North Carolina. Currently there are 100 counties in North Carolina, the study appears to encompass 97 of those counties.

1.1.2 Background

There are 3 classes of variables provided to us: 1. Labor market variables 2. Criminal justice variables 3. Demographic variables

Both labor market and criminal justice strategies should be relevant in influencing (causing causal behavior on) crime rate.

In our study, we are given variables that reflect both these factors: **Variables reflecting labor markets**

10 taxpc - tax revenue per capita 15 wcon - weekly wage, construction 16 wtuc - wkly wge, trns, util, commun 17 wtrd - wkly wge, whlesle, retail trade 18 wfir - wkly wge, fin, ins, real est 19 wser - wkly wge, service industry 20 wmfg - wkly wge, manufacturing 21 wfed - wkly wge, fed employees 22 wsta - wkly wge, state employees 23 wloc - wkly wge, local gov emps

Variables reflecting criminal justice system

4 prbarr - 'probability' of arrest 5 prbconv - 'probability' of conviction 6 prbpris - 'probability' of prison sentence 7 avgsen - avg. sentence, days 8 polpc - police per capita 24 mix - offense mix: face-to-face/other

There is a third class of variables that represents the demographic characteristics of the data. We expect some of these variables to have a causal effect on crime rate (say pctymle or pctmin80), but others to not intrinsically have causal effects on crime rate (say west, central, county).

These variables can also possibly influence the "labor market" and "criminal justice system" variables:

Variables reflecting demographic characteristics

1 county - county identifier 9 density - people per sq. mile 11 west - =1 if in western N.C. 12 central - =1 if in central N.C. 13 urban - =1 if in SMSA 14 pctmin80 - perc. minority, 1980 25 pctymle - percent young male

1.2 2. Research Question

We are interested in understanding the relationship of crime that may be caused by these 3 classes of variables (or their interactions). We are looking to see if any one class of variables is dominant in deciding crime rate in the area. Based on the variables we do identify, there may be different sets of implications for the local government from a policy perspective. For example, if we determine "criminal justice" variables are primarily responsible for crime rate, then our recommendations may be to change certain pre and post conviction policies to curb crime. If, on the other hand the "labor market" variables are primarily causing crime, that may prompt government response to create more jobs in a certain industry, for example. Its also possible that more than 1 class of variables may be at play here.

1.3 3. Operationalization of Model Specification

Our chosen target variable is "crmrate", and we're trying to understand what affects this variable (or a transform of this variable).

Note that we are choosing crime rate as our target variable with the intention to predict crime rate from the data. This is consistent with previous research that has been done

https://www.amherst.edu/media/view/121570/original/CornwellTrumbullCrime%2BElasticities.pdf

We could have argued to choose prbarr, prbconv, or prbpris as our target variables and chosen crmrate as our predictor. This is because we **CANNOT assume the direction of causality between crime rate and eg. probability of getting arrested (prbarr)**. We can argue that an increased crime rate causes an increase in probability of arrests, just as well as arguing that prbarr causes crime rate. For now, we will choose some transformation of crime rate as our target variable.

For independent variables, we will do a thorough EDA on all classes of variables to understand which variables to include in the model. We will also consider certain interaction variables to create intuitive metrics that may inform crime.

1.4 4. Data Loading and Cleaning

[5]:
```
install.packages("dplyr")
library(dplyr)
install.packages("car")
library(car)
install.packages("lmtest")
library(lmtest)
install.packages("sandwich")
library(sandwich)
install.packages("corrplot")
library(corrplot)
install.packages("stargazer")
library(stargazer)
install.packages("curl")
library(curl)
```

```
install.packages("data.table")
library(data.table)
install.packages("haven")
library(haven)
install.packages("readxl")
library(readxl)
install.packages("ggplot2")
library(ggplot2)
install.packages("corrplot")
library(corrplot)
install.packages("GGally")
library(GGally)
# set standard height and width for images displayed
options(repr.plot.width=5, repr.plot.height=5)
```

```
Warning message:
"package 'dplyr' is in use and will not be installed"Warning message:
"package 'car' is in use and will not be installed"Warning message:
"package 'lmtest' is in use and will not be installed"Warning message:
"package 'sandwich' is in use and will not be installed"Warning message:
"package 'corrplot' is in use and will not be installed"Warning message:
"package 'stargazer' is in use and will not be installed"Warning message:
"package 'curl' is in use and will not be installed"Warning message:
"package 'data.table' is in use and will not be installed"Warning message:
"package 'haven' is in use and will not be installed"Warning message:
"package 'readxl' is in use and will not be installed"Warning message:
"package 'ggplot2' is in use and will not be installed"Warning message:
"package 'corrplot' is in use and will not be installed"Warning message:
```

[2]:
```
get_data = function() {
    wd = getwd()
    return (read.csv(paste(wd, "/", "crime_v2.csv", sep="")))
}
```

[3]:
```
data = get_data()
# head(data)
```

[4]:
```
toString(sapply(data, class))
```

'integer, integer, numeric, numeric, factor, numeric, numeric, numeric, numeric, numeric, integer, integer, integer, numeric, numeric, numeric, numeric, numeric, numeric, numeric, numeric, numeric, numeric, numeric, numeric'

We notice there is a factor column for prbconv, we convert this to numeric, as has been done with prbarr and prbpris. We also observe that there are few missing values for prbconv. These will be removed later

[6]:
```
data$prbconv = as.numeric(as.character(data$prbconv))
```

```
Warning message in eval(expr, envir, enclos):
```

Notice that the 6 rows of NA values above are coming because prbconv is NA in those 6 rows

[7]:
```
# we observe all values have NA in them
# remove data where there are NA in rows
data = data[complete.cases(data), ]
```

[8]:
```
dim(data)
```

1. 91 2. 25

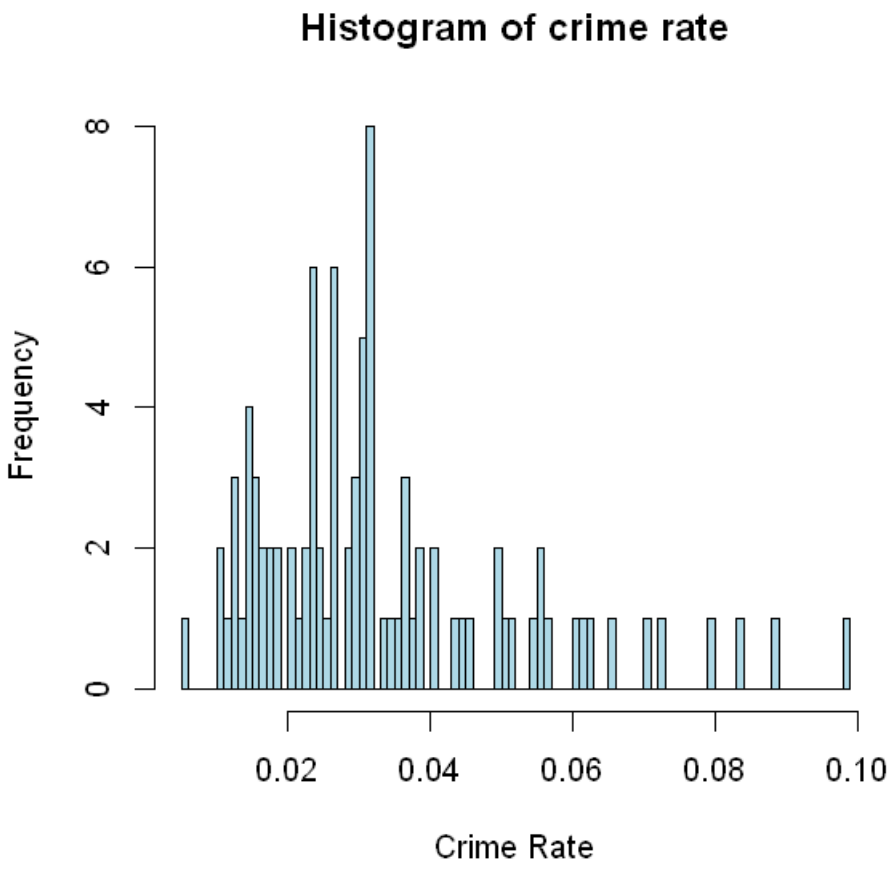
We observe we have removed the 6 rows and now we have 91 rows.

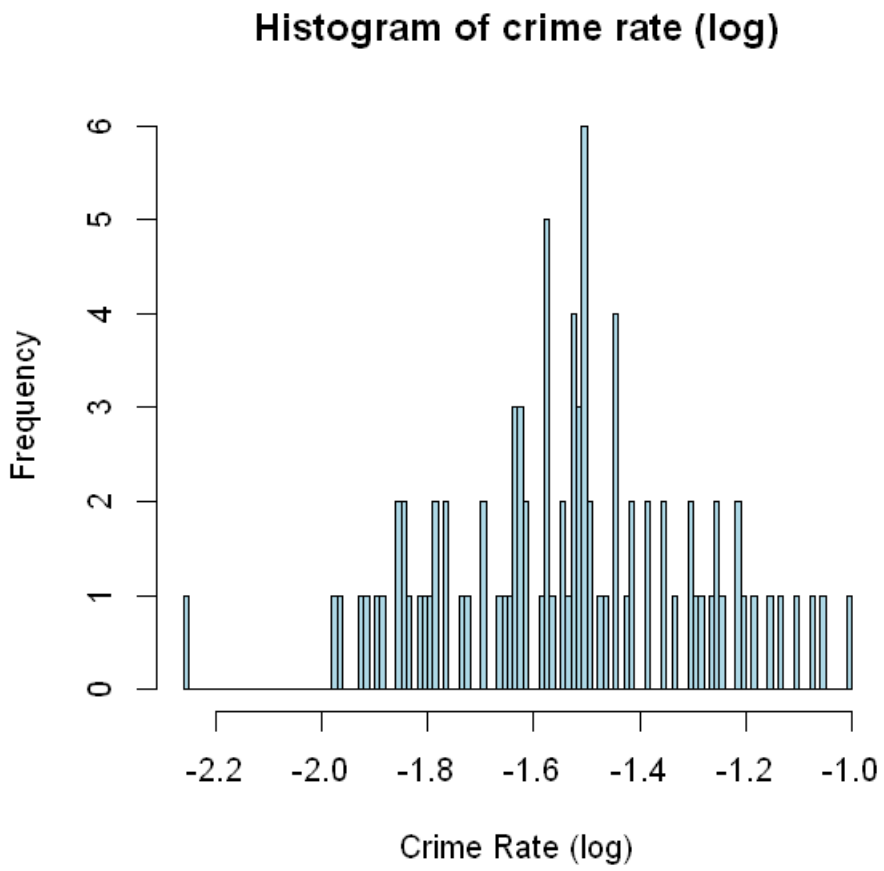
1.4.1 Exploratory Data Analysis

1.4.2 EDA Target Variable

[9]:
```
summary(data$crmrte)
hist(data$crmrte, breaks=100, main = "Histogram of crime rate", xlab = "Crime␣
 ↪Rate", col = "lightblue", border = "black")
hist(log10(data$crmrte), breaks=100, main = "Histogram of crime rate (log)",␣
 ↪xlab = "Crime Rate (log)", col = "lightblue", border = "black")
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.005533 0.020927 0.029986 0.033400 0.039642 0.098966
```


Histogram of crime rate
Frequency
0
2
4
6
8
0.02
0.04
0.06
0.08
0.10
Crime Rate



Note that if we take a log version of the crime rate, we get a nice normally distributed curve. Hence, we will consider our target variable to be log10(crmrte)

[10]:
```
data$y = log10(data$crmrte)
```

1.4.3 EDA Geographical Variables (Demographic)

One thing to consider would be to consider the effects of geography on crime rate.

There are 3 variables which help us categorize geography - west, central and urban. These are all dummy variables that represent what region characteristics the data point has. Hence, one variable of interest could be an interaction variable:

geo = west x central x urban Note that because we include this new interaction variable, we have to include all lower order interactions as well.

[11]:
```
data$geo = data$west * data$central * data$urban
data$west_central = data$west * data$central
data$west_urban = data$west * data$urban
```

```
data$central_urban = data$central * data$urban

# also include a variable to capture non urban, non west, non central region␣
↪(west=0, central=0, urban=0)
geo_arr = data$west + data$central + data$urban
# hist(geo_arr, , breaks=100, main = "Histogram of Urban, Central, West Combo",␣
↪xlab = "Sum(Urban, West, Central)", col = "lightblue", border = "black")
# notice there are over 30 counties that are non urban, non west, non central
data$non_urban_west_central = ifelse(geo_arr >= 1, 0, 1)
```

[12]:
```
# plot(data$urban, data$density, main = "Urban vs Population Density", xlab =␣
↪"Urban", ylab = "Population Density" )
```

We don't have a hypothesis around geography, but one can make an argument that urban areas are more susceptible for crime, given that they may be more densely populated.

Also, we notice that there are no regions that are urban, west and central (see graph above for sum(urban, west, central)). There are also many areas (over 30 counties) that are neither urban, nor west nor central. This gives a basic idea of geographical characteristics of the region

[13]:
```
# Analysis of interaction term "geo"
paste("Urban:", sum(data$urban))
paste("West:", sum(data$west))
paste("Central:", sum(data$central))

paste("Central + West:", sum(data$central * data$west))
paste("Central + Urban:", sum(data$central * data$urban))
paste("Urban + West:", sum(data$urban * data$west))

paste("Urban + West + Central:", sum(data$urban * data$west * data$central))
paste("NON Urban + NON West + NON Central:", sum(data$non_urban_west_central))
```

'Urban: 8'
'West: 23'
'Central: 34'
'Central + West: 1'
'Central + Urban: 5'
'Urban + West: 1'
'Urban + West + Central: 0'
'NON Urban + NON West + NON Central: 33'

For now, we will include the above variables in our model to observe if there are any special effects on crime based on geo

1.4.4 EDA Population Variables (Demographic)

We decide to add 2 interaction variables: 1. density x pctymle - this gives the density of young males in the county (density_ymle) 2. density x pctmin80 - this gives the density of minority in the county (density_min80)

These seem to be a good representation of the density of a population in the county, which is a more intuitive way to think about demography (people / unit area), rather than thinking in percentage terms only. Hence, we will add these variables into our discussion as well.

Note that this assumes that the percentage variables are uniformly distributed over the county and that density x pct will give the density of the population of interest in the county.

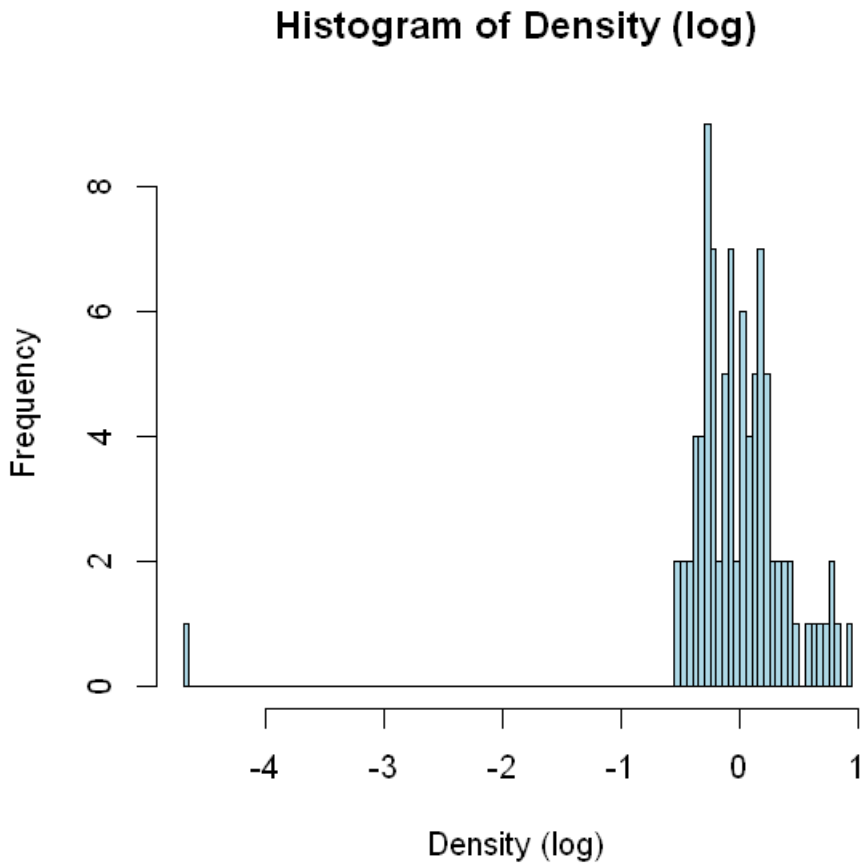
Also note that pctymle is already given in fractional terms (between 0 and 1), whereas pctmin80 is given in percentage (between 0 and 100). For sake of uniformity, we will convert pctmin80 to 0-1 scale

[14]:
```
data$pctmin80 = data$pctmin80 * 1.0 / 100
# summary(data$pctmin80)
```

[15]:
```
data$density_min80 = data$density * data$pctmin80
data$density_ymle = data$density * data$pctymle
```

[16]:
```
summary(data$density)
# hist(data$density, , breaks=100, main = "Histogram of Density", xlab =␣
 ↪"Density", col = "lightblue", border = "black")
hist(log10(data$density), , breaks=100, main = "Histogram of Density (log)",␣
 ↪xlab = "Density (log)", col = "lightblue", border = "black")
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00002 0.54741 0.96226 1.42884 1.56824 8.82765
```



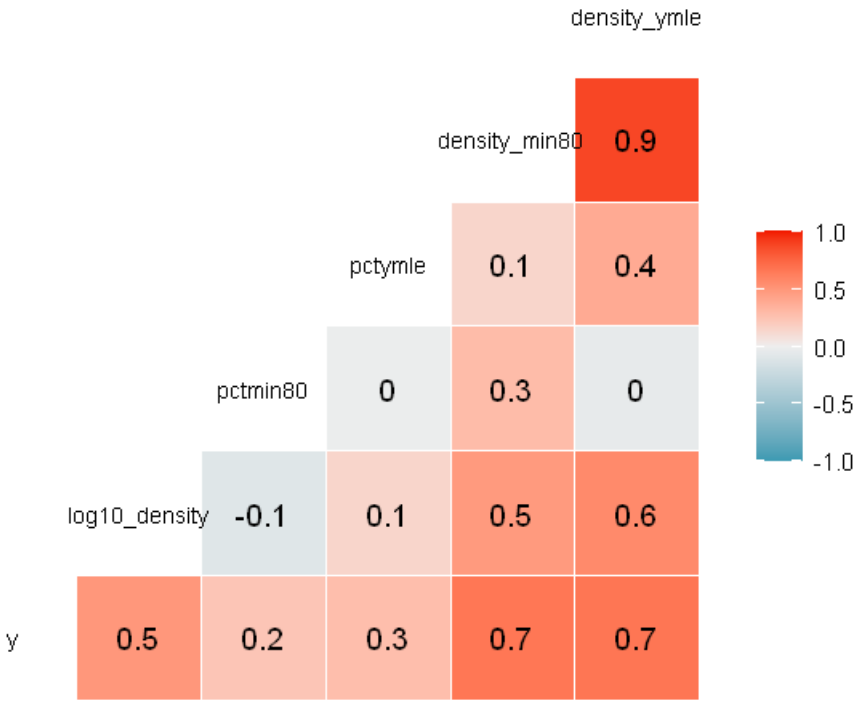
Taking the log(Density) makes the data much more normal and is a better choice for our density variable.

Doing a histogram of pctmin80 and pctymle (or their log transformations) - didnt reveal any interesting normal distributions. We notice that the values for the percentages lie within expected range (0 to 1). So we will continue to explore further.

Next lets do a correlation analysis on all the variables of population identified so far.

[17]:
```
data$log10_density = log10(data$density)
```

[18]:
```
ggcorr(data[, c("y","log10_density", "pctmin80", "pctymle", "density_min80",␣
↪"density_ymle")], size = 3, label=TRUE)
```



We observe strong positive correlation between several population variables Here y is log10(crmrte)

1. High correlation between crime rate and density_ymle

2. High correlation between crime rate and density_min80
3. High correlation between crime rate and log10_density

We will include these variables in our specifications with the following expected causal effects: 1. If density is high, then crime rate is expected to be high. This is simply because a larger set of people in one area is bound to create more conflict and crime in the community. 2. If percentage of young males are high in a region, we expect more crime. This is because younger people are easier to influence and may engage more in petty crimes. Although this is not explicitly shown with large correlation above, its worth investigating further.

There may be a concern including both density_min80 and density_ymle because they have high correlation (0.88) - leading to possible multicollinearity. This also shows that the density of younger male populations is highly correlated across counties with density of minority populations. Keep in mind, pctymle and pctmin80 are hardly correlated with each other. But their density equivalents (density_min80 and density_ymle) are highly correlated, which is an interesting observation. This could be because of the underlying common density variable in the interaction. For now, we will keep density_ymale in our model, as a measure of how many young males per unit area exist in each county. We will NOT include density_min80 because we are concerned about multicollinearity. To represent the minority percentage variable, we keep pctmin80 in the model.

For now, we will include "log10(density)", "pctmin80" and "density_ymle" variables in the model to observe their effects

1.4.5 EDA Criminal Justice Variables

It is worthwhile to capture any relationships between these variables: prbarr, prbconv, prbpris.

This will help us determine if there is any multicollinearity between these variables. We would suspect these to be correlated

[19]:
```
# summary(data$prbpris)
# hist((data$prbpris), breaks=100, main = "Histogram of prbpris", xlab =
 ↪"prbpris", col = "lightblue", border = "black")
```

[20]:
```
# summary(data$prbarr)
# hist((data$prbarr), breaks=100, main = "Histogram of prbarr", xlab =
 ↪"prbarr", col = "lightblue", border = "black")
```

Notice that there is an outlier row with a probability higher than 1, which is not possible. We will scale our data to make probabilities <=1

[21]:
```
data$prbarr = data$prbarr / max(data$prbarr)
```

[22]:
```
# summary(data$prbconv)
# hist((data$prbconv), breaks=100, main = "Histogram of prbconv", xlab =
 ↪"prbconv", col = "lightblue", border = "black")
```

Notice that there are a few rows where prbconv is > 1 even though it is supposed to be a probability. We will scale our data to make probabilities <=1

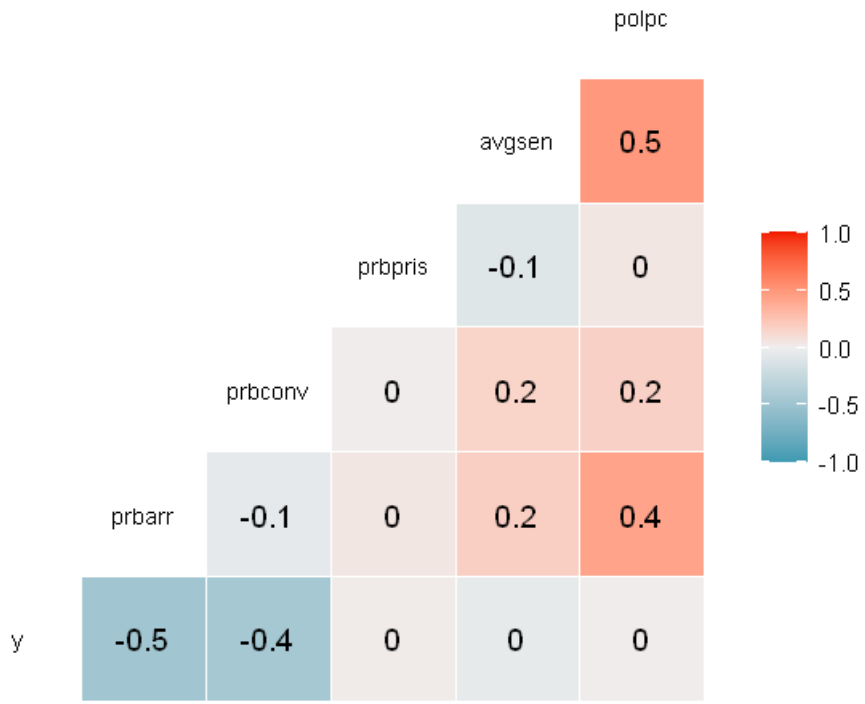
[23]:
```
data$prbconv = data$prbconv / max(data$prbconv)
```

Its important to note that log(data$prbconv) results in a slightly more normal representation of the prbconv variable, however, we lose interpretation of the variable.

We are not sure what log(prbconv) would mean in terms of explanability, so we DO NOT include it in the model.

Now that we have the data in the right format, we find correlation between the 3 probabilities, along with average sentence and police per capita

[24]:
```
ggcorr(data[, c("y", "prbarr", "prbconv", "prbpris", "avgsen", "polpc")], size␣
↪= 3, label=TRUE)
```



Correlation Analysis

1. Interestingly, there is little correlation between the probability variables
2. There is positive correlation between polpc and avgsen
3. There is positive correlation between polpc and prbarr
4. There is negative correlation between crime rate and (prbarr, prbconv)

For now, we will consider all 3 probability variables in the model

Its interesting to note that prbarr and prbconv influence crime rate, but prbpris has little to do with crime rate. This implies that "before the criminal is imprisoned" has high associativity to crime rate.

We expect that as any of probabilities of arrest or conviction or prison sentence go up, crime rate should come down (people would be less risk taking to commit crime if they knew they would be likely arrested / convicted / imprisoned).

[25]:
```
# Finally there's nothing special about the avgsen and polpc variables.

# They have values within range and don't display special distributions
# summary(data$avgsen)
# hist(data$avgsen, breaks=100, main = "Histogram of avgsen", xlab = "avgsen",␣
 ↪col = "lightblue", border = "black")

# Taking log10(avgsen) makes the histogram look slightly more normal, but we␣
 ↪will use avgsen as is in the model for now.
# hist(log10(data$avgsen), breaks=100, main = "Histogram of avgsen", xlab =␣
 ↪"avgsen", col = "lightblue", border = "black")

# summary(data$polpc)
# hist(data$polpc, breaks=100, main = "Histogram of polpc", xlab = "polpc", col␣
 ↪= "lightblue", border = "black")
```

There seem to be a few suspicious far right outliers for police per capita and average sentence days

We will deal with them as appropriate when evaluating the cook's distance for potential points of influence.

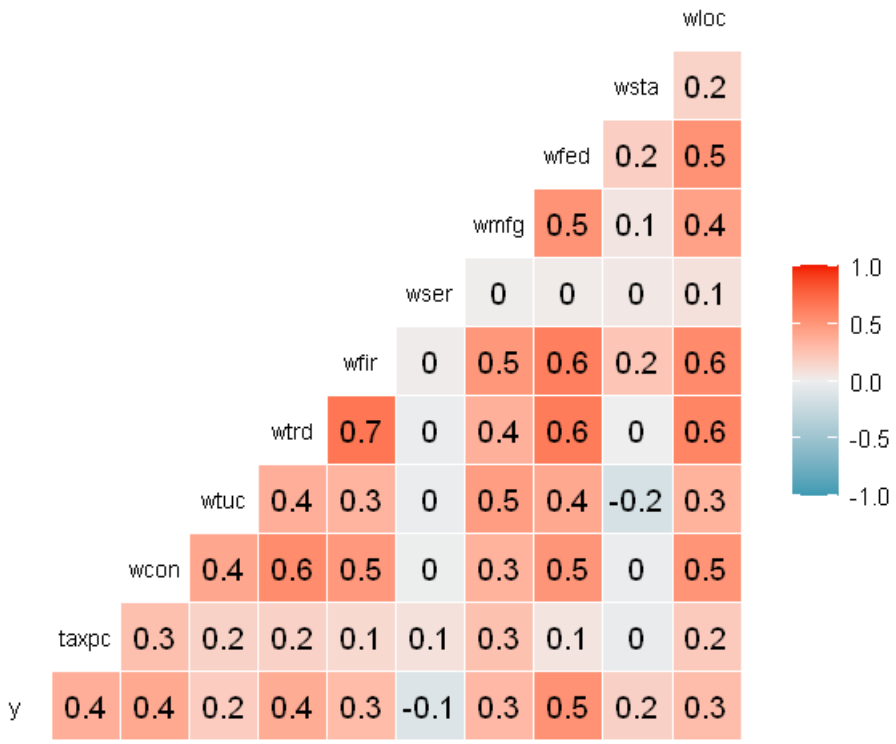
We cannot make a judgement for the causal effect of polpc (police per capita). If this increases, we can make an argument that crimes may go down as there is "more policing" per citizen. We can also say, if polpc goes up, then crime rate goes up (police can find trivial reasons to arrest citizens - overpolicing).

We can however, likely say, if avgsen (average sentence in days) goes up, then convicts may become may learn their lessons and not commit crimes the next time around, so crime rate decreases.

For now, we will include both variables, as is in the model specification

1.4.6 EDA Labor Market Variables

[26]:
```
ggcorr(data[, c("y", "taxpc", "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg",␣
 ↪"wfed", "wsta", "wloc")], size = 3, label=TRUE)
```



As expected, due to labor and market interdependencies, a lot of the wages of workers in the city are correlated with others. Note that wages are positively correlated across all industries.

Note that while there is strong positive correlation (> 0.4), there is no perfect multicollinearity amongst these variables

Tax revenues are hardly correlated with any of the wages, this is a little surprising! Although we do see weak correlation between taxes and certain industries.

Crime rate is positively correlated taxes, as well as few other industries (wcon, wfrd, wfed)

From this article: https://www.amherst.edu/media/view/121570/original/CornwellTrumbullCrime%2BEl

We find that these variables of wage represent **average** weekly wages of the industry

One variable of interest could be the total weekly wages:

wtotal = wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc

The argument is that if the total average weekly wages are high for a county, then the crime rate will be low (as the county is prosperous, so robberies and money related crimes would be fewer)

One may wonder if its valid to add wages and include this in the model. There are a few arguments for doing this: 1. Adding all wages shouldnt affect the regression because its the same effect as a linear combination of the betas of the individual industrial wages. In other words, the coefficient of wtotal is an aggregated version of the effect of each of the coefficients of the wage

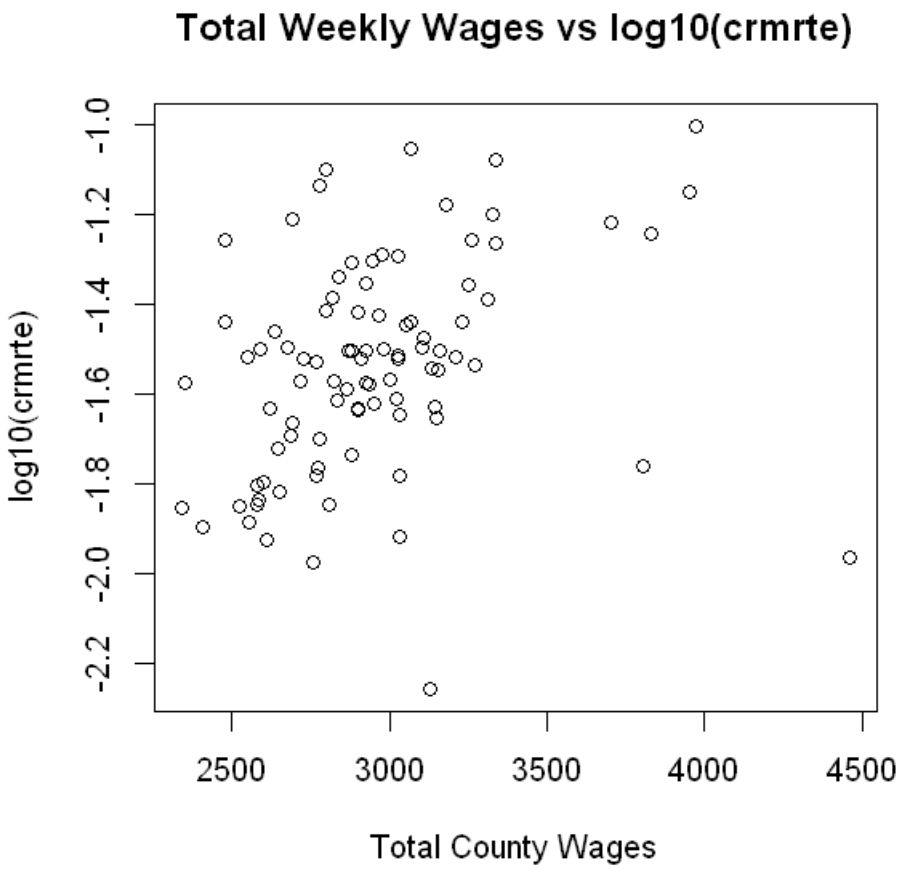
variables. 2. Adding all wages gives an idea of the overall economic condition of the county, which would affect crime rate of the county on aggregate 3. **Even if wages in 1 county were not distributed evenly amongst sectors (on average), but the total county is overall wealthy, that may be an argument to influence the policy makers to tax the rich!** Higher wage workers can be taxed more to provide services and benefit to the poorer wage workers in the county. If poor wage earners are now taken care of with the wealth and services from the government, there may be lesser incentive for crime in the county. This scenario is admittedly a little imaginative, but it is a plausible policy suggestion that could be put forward if wtotal is a significant variable in the model!

There are a few caveats with this variable though: 1. Total average weekly wage in a county, while succinct, may be a controversial variable to include because individual wages are aggregated into a single number and may be hiding the effects of each individual industry. For example, consider wcon, wfir and wmfg variables. If in county1, the values are 100, 200, 300 and in county2, the values are 1, 2, 597. wtotal is the same for both counties, however, the breakdown shows that the workforce in county2 is on average poorer than county1, and thus susceptible to crime.

[27]:
```
data$wtotal = data$wcon + data$wtuc + data$wtrd + data$wfir + data$wser +␣
↪data$wmfg + data$wfed + data$wsta + data$wloc
```

[28]:
```
plot(data$wtotal, data$y, main = "Total Weekly Wages vs log10(crmrte)", xlab =␣
↪"Total County Wages", ylab = "log10(crmrte)")
cor(data$wtotal, data$y)
```

0.311251670277



Total weekly wages is not unreasonably skewed in distribution (except for 5 outliers). More or less, each wage variable follows a normal distribution (as expected for population wages). **Hence, it is reasonable to use wtotal in our model.**

1.5 5. Model Building

Our target variable is log10(crmrate) - which is a scaled version of the log of crime rate. This gives us percentage increase of crime rate

We will build 3 model specifications that consist of one or more **classes** of variables:

1. Model1: Criminal Justice
2. Model2: Select variables our team thinks are important causes
3. Model3: Maximalist approach model with backward elimination / forward selection

We will use Adjusted R-Squared, AIC, BIC for criteria of model fit

[29]:
```
dim(data)
```

1. 91 2. 35

1.5.1 Demographic Variables Considered

Geo variables

1. geo
2. west_central
3. west_urban
4. central_urban
5. central
6. west
7. urban
8. non_urban_west_central

Population variables

9. log10_density
10. density_ymle (possibly, pctymle)
11. pctmin80

1.5.2 Criminal Justice Variables Considered

All variables listed in the criminal justice category are considered without transformation 1. prbarr 2. prbconv 3. prbpris 4. avgsen 5. polpc

1.5.3 Labor Market Variables Considered

1. wtotal
2. taxpc
3. Optionally, each industy's wage

Note that we have already run our models before and determined that a few rows introduce very strong outlier tendencies, and have very strong influence (large cook's distance). **We will eliminate these rows from consideration** because these rows have caused coefficients that were originally significant to become insignificant in on our previous model runs.

[30]:
```
outlier_row = 79
data[outlier_row, ]
dim(data)
```

A data.frame: 1 Œ 35

	county <int>	year <int>	crmrte <dbl>	prbarr <dbl>	prbconv <dbl>	prbpris <dbl>	avgsen <dbl>	polpc <dbl>
79	173	87	0.0139937	0.4862317	0.154567	0.15	6.64	0.0031637

1. 91 2. 35

[31]:
```
data = data[-outlier_row, ]
dim(data)
```

1. 90 2. 35

[32]:
```
outlier_row = 83
data[outlier_row, ]
dim(data)
```

A data.frame: 1 Œ 35

	county <int>	year <int>	crmrte <dbl>	prbarr <dbl>	prbconv <dbl>	prbpris <dbl>	avgsen <dbl>	polpc <dbl>
84	185	87	0.0108703	0.1789937	1	0.442857	5.38	0.001222

1. 90 2. 35

[33]:
```
data = data[-outlier_row, ]
dim(data)
```

1. 89 2. 35

[34]:
```
outlier_row = 51
data[outlier_row, ]
dim(data)
```

A data.frame: 1 Œ 35

	county <int>	year <int>	crmrte <dbl>	prbarr <dbl>	prbconv <dbl>	prbpris <dbl>	avgsen <dbl>	polpc <dbl>
51	115	87	0.0055332	1	0.7071435	0.5	20.7	0.00905433

1. 89 2. 35

[35]:
```
data = data[-outlier_row, ]
dim(data)
```

1. 88 2. 35

1.5.4 Model1 : Base Model

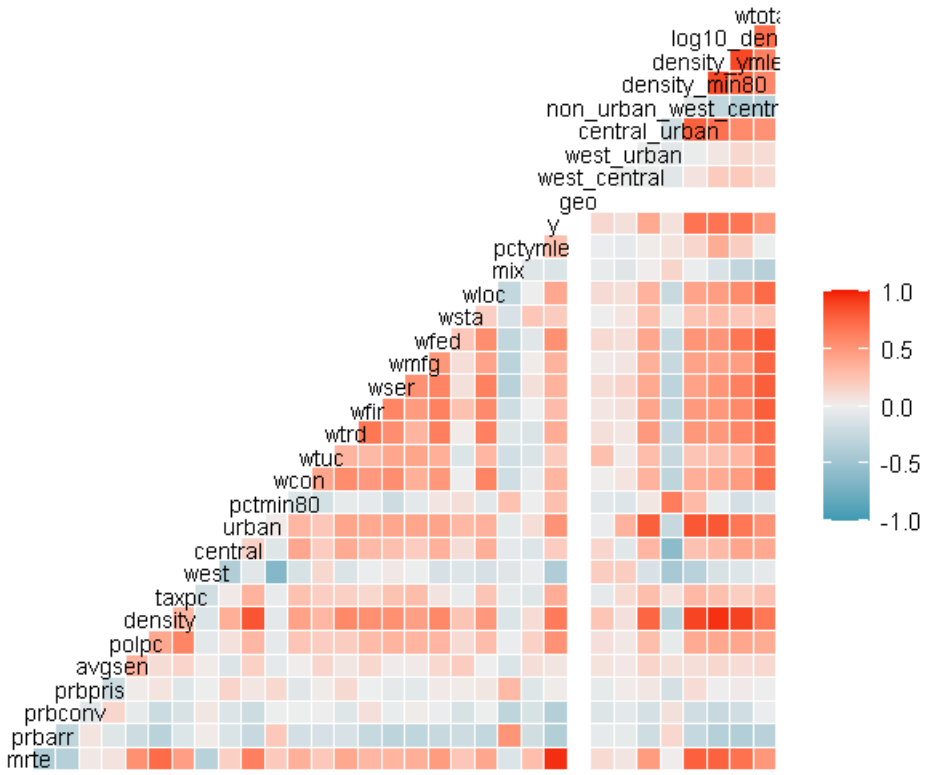
Model1 selects the following variables - log10(density), prbarr, prbconv, pctmin80, polpc

These are selected because these variables have the highest correlation with log10(crmrate) based on our correlation plots.

This gives us a first pass around modeling the data such that we can create a model on variables that show most change with respect to crime. Moreover, most of the variables that did show correlation also intuitively seem to make sense to include in the model.

[36]:
```
ggcorr(data[, 3:35], size = 3)
```

```
Warning message in cor(data, use = method[1], method = method[2]):
```



[37]:
```
model1 = lm(y ~ log10_density + prbarr + prbconv + pctmin80 + polpc, data =␣
↪data)
```

We'll compare this model with our balanced model and maximalist model later. This is our baseline model.

1.5.5 Model2: Our Selected Model (Balanced Approach)

We choose a balanced model, which we believe gives us the right variables for predicting crime rate, based on our base model (model1), along with the EDA done so far.

We will analyze a modified version of this model in detail to test for MLR assumptions and for model performance.

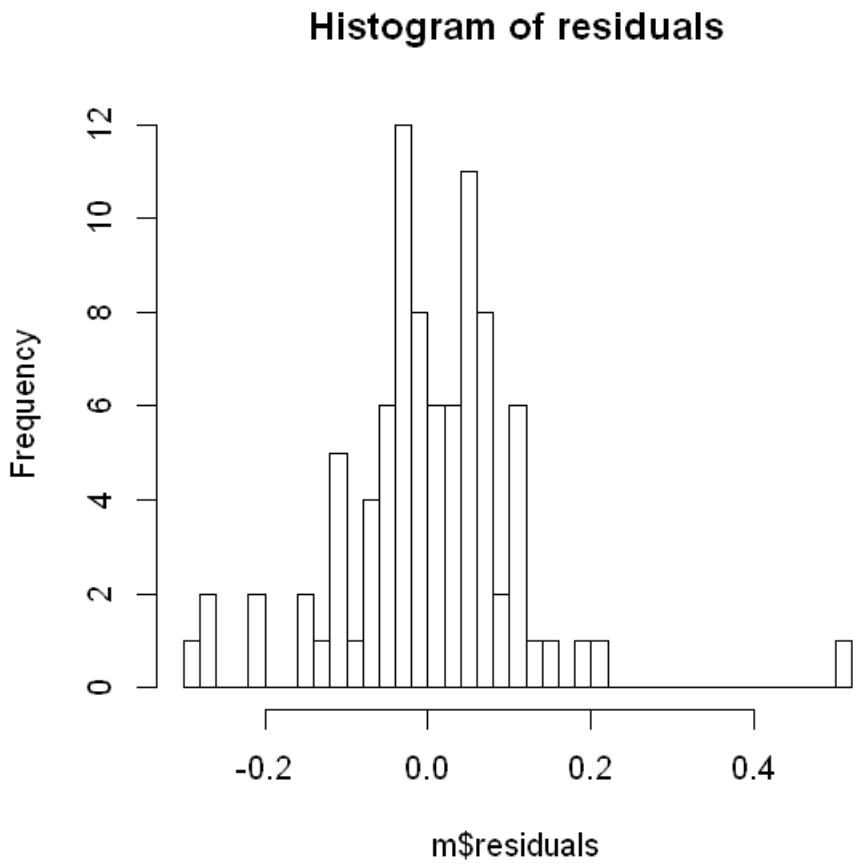
[38]:
```
model2 = lm(y ~ prbarr + prbconv + polpc + log10_density + pctmin80 + pctymle +␣
↪wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc, data=data)
```

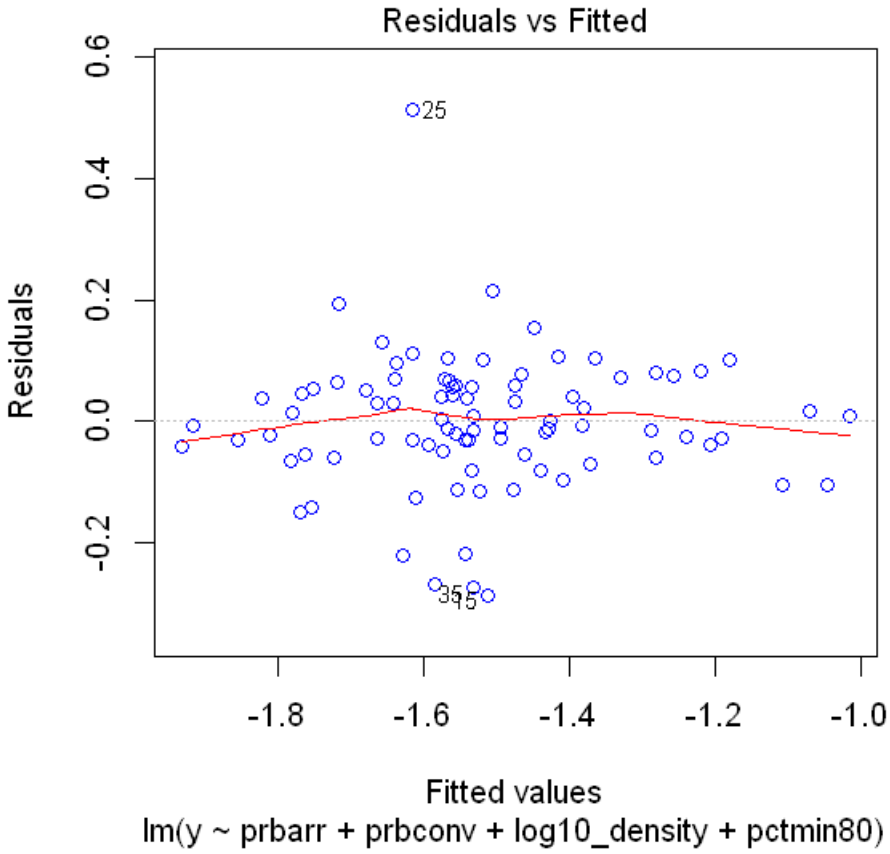
[42]:
```
# set model to be evaluated here

model = model2
```

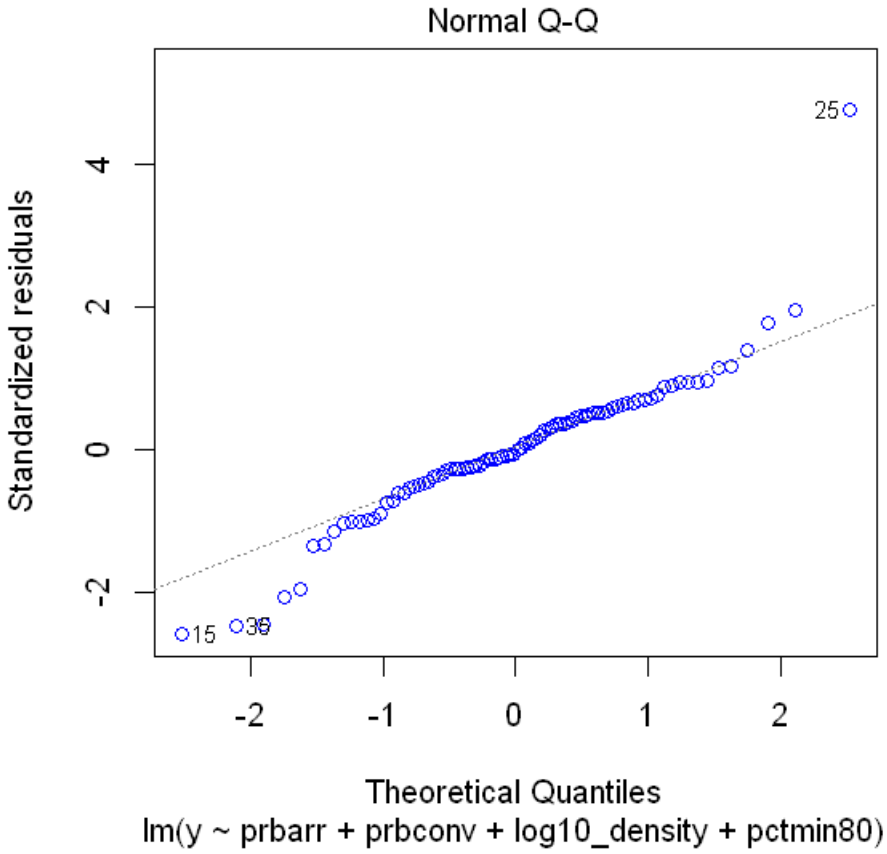
[43]:
```
# We have kept all models linear in beta coefficients (MLR 1 satisfied for all␣
 ↪models)
# Note that we assume data has been collected as IID (MLR 3 for all models)
evaluate_model_shapiro = function(m) {
    paste("Shapiro test on residuals - test for normality (MLR 6)")
    # null hypothesis is residuals are normally distributed
    shapiro.test(m$residuals)
}

evaluate_model_bp = function(m) {
    paste("Breusch Pagan test - test for heteroskedasticity (MLR 5)")
    # null hypothesis is homoskedasticity, alt is presence of␣
 ↪heteroskedasticity
    bptest(m)
}


evaluate_model_plots = function(m) {

    paste("Plotting residuals - test for normality of errors (MLR 6)")
    hist(m$residuals, breaks=50, main = "Histogram of residuals")

    paste("Diagnostic plots (MLR 4,5 - Residual vs fitted), (MLR6 - QQ Plot),␣
 ↪(MLR5 - Std. Residuals), (Cooks Dist - Outliers test)")
    #par(mfrow=c(3,3))
    plot(m, which=1:6, col = 'blue')

}

summary(model$residuals)
evaluate_model_shapiro(model)
evaluate_model_bp(model)
evaluate_model_plots(model)
```

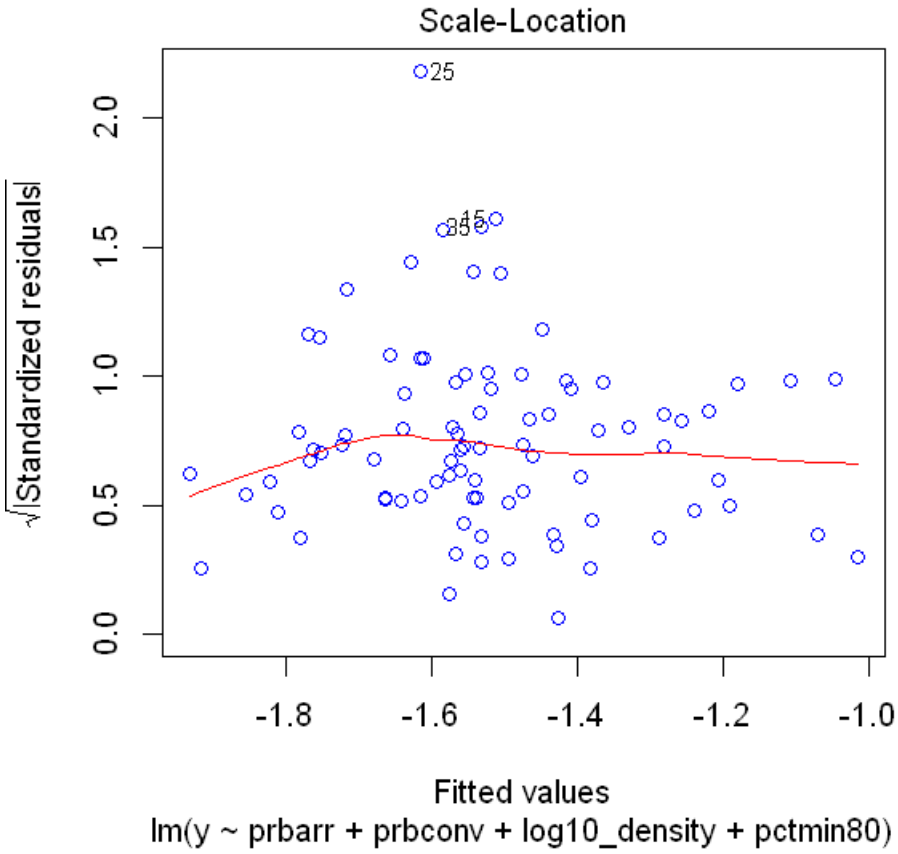
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-0.285977 -0.050699 -0.003276  0.000000  0.058572  0.512541


Shapiro-Wilk normality test

data:  m$residuals
W = 0.91463, p-value = 2.373e-05



studentized Breusch-Pagan test

data:  m
```

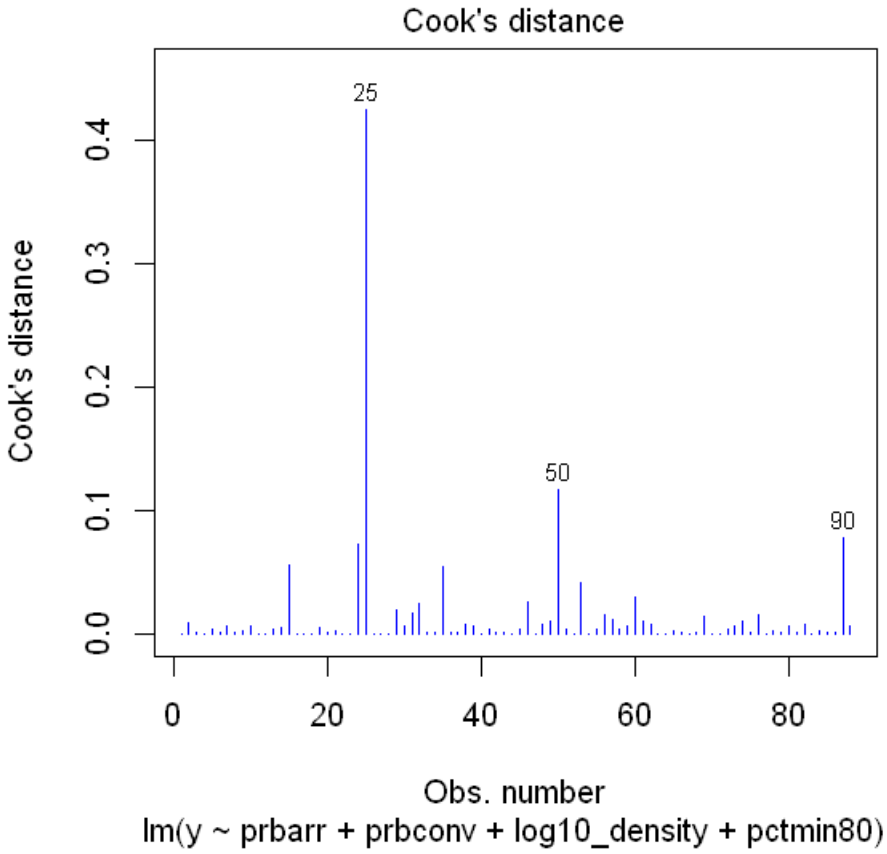
```
BP = 9.8763, df = 4, p-value = 0.04256
```

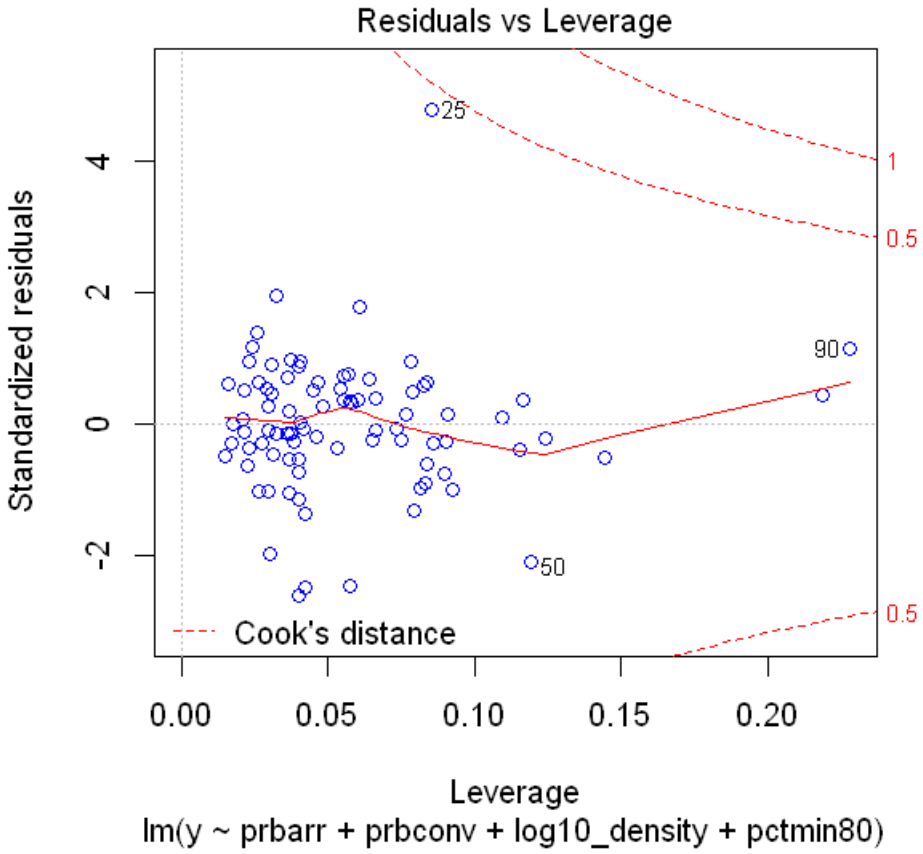


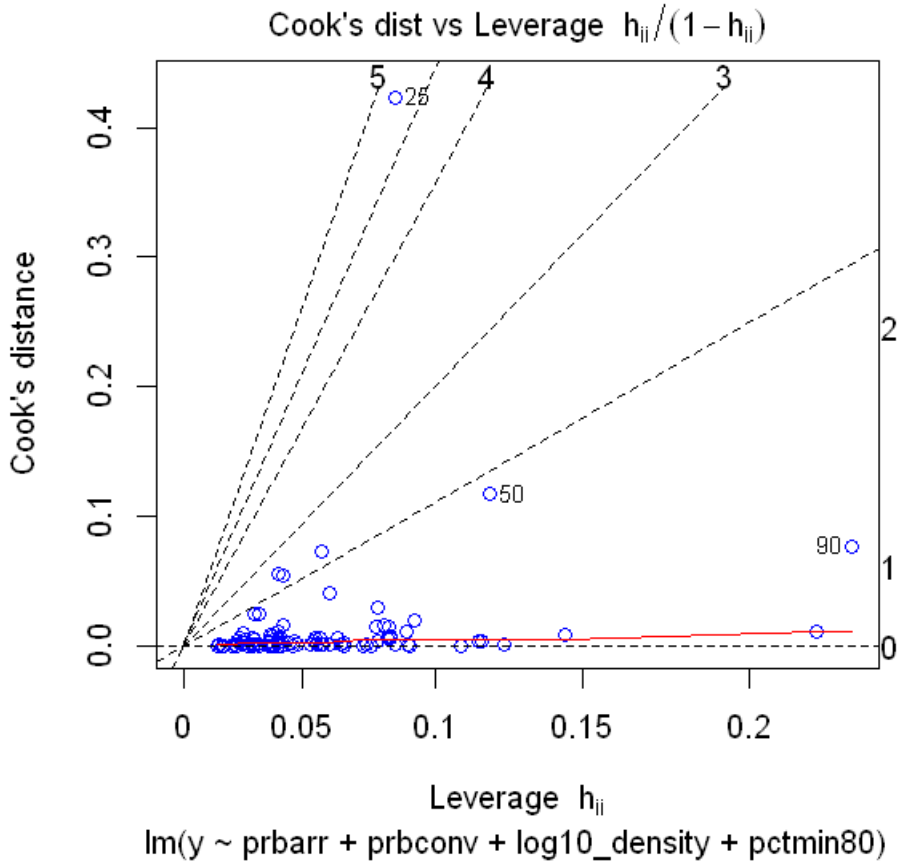

Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
-0.2
25
35
15
-1.8
-1.6
-1.4
-1.2
-1.0
Fitted values
lm(y ~ prbarr + prbconv + log10_density + pctmin80)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(y ~ prbarr + prbconv + log10_density + pctmin80)
25
15
-2
-1
0
1
2
4


Scale-Location
25
15
35
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
-1.8
-1.6
-1.4
-1.2
-1.0
Fitted values
lm(y ~ prbarr + prbconv + log10_density + pctmin80)


Cook's distance
25
50
90
Cook's distance
0.0
0.1
0.2
0.3
0.4
0
20
40
60
80
Obs. number
lm(y ~ prbarr + prbconv + log10_density + pctmin80)


Residuals vs Leverage
Standardized residuals
4
2
0
-2
25
90
50
1
0.5
0.5
Cook's distance
0.00
0.05
0.10
0.15
0.20
Leverage
lm(y ~ prbarr + prbconv + log10_density + pctmin80)



[44]:
```
# Always use heteroskedasticity robust std errors for beta values
# Get the coefficients for the variables in the model
coeftest(model, vcov = vcovHC)
summary(model)
AIC(model)
BIC(model)
vif(model)
```

```
t test of coefficients:

               Estimate Std. Error  t value  Pr(>|t|)
(Intercept)   -1.315140   0.092179 -14.2672 < 2.2e-16 ***
prbarr        -0.850927   0.174202  -4.8847 4.968e-06 ***
prbconv       -0.570907   0.133971  -4.2614 5.335e-05 ***
log10_density  0.364965   0.059863   6.0967 3.277e-08 ***
pctmin80       0.580710   0.076562   7.5849 4.338e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Call:
lm(formula = y ~ prbarr + prbconv + log10_density + pctmin80,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28598 -0.05070 -0.00328  0.05857  0.51254

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.31514    0.05533 -23.771  < 2e-16 ***
prbarr        -0.85093    0.14789  -5.754 1.42e-07 ***
prbconv       -0.57091    0.09559  -5.973 5.58e-08 ***
log10_density  0.36497    0.04205   8.678 2.87e-13 ***
pctmin80       0.58071    0.07488   7.755 1.99e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1125 on 83 degrees of freedom
Multiple R-squared:  0.7525,Adjusted R-squared:  0.7406
F-statistic: 63.09 on 4 and 83 DF,  p-value: < 2.2e-16
```

-127.948938886964

-113.084918000095

prbarr 1.4215104373298 **prbconv** 1.2342815575076 **log10_density** 1.31377454357741 **pctmin80** 1.06176191643656

We see that from our original model2 (including wage variables) there are some wage variables that are showing partial significance in our model.

Hence, just to be sure of their "significance" in the model, we conduct a joint significance test (wald test)

[39]:
```
linearHypothesis(model2, c("wfir = 0", "wser = 0", "wfed = 0", "wsta = 0"),␣
↪vcov = vcovHC)
```

A anova: 2 Œ 4

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	76	NA	NA	NA
2	72	4	2.397085	0.05806614

[40]:
```
linearHypothesis(model2, c("wcon = 0", "wtuc = 0", "wtrd = 0", "wfir = 0",␣
↪"wser = 0", "wmfg = 0", "wfed = 0", "wsta = 0", "wloc = 0"), vcov = vcovHC)
```

A anova: 2 Œ 4

	Res.Df <dbl>	Df <dbl>	F <dbl>	Pr(>F) <dbl>
1	81	NA	NA	NA
2	72	9	1.274639	0.2656056

We observe that these variables are hardly passing the wald test (linearHypothesis) for joint significance at the 5% level.

Hence, we fail to reject the null (where each beta value = 0). This suggests that the wage variables are indeed not contributing to the model.

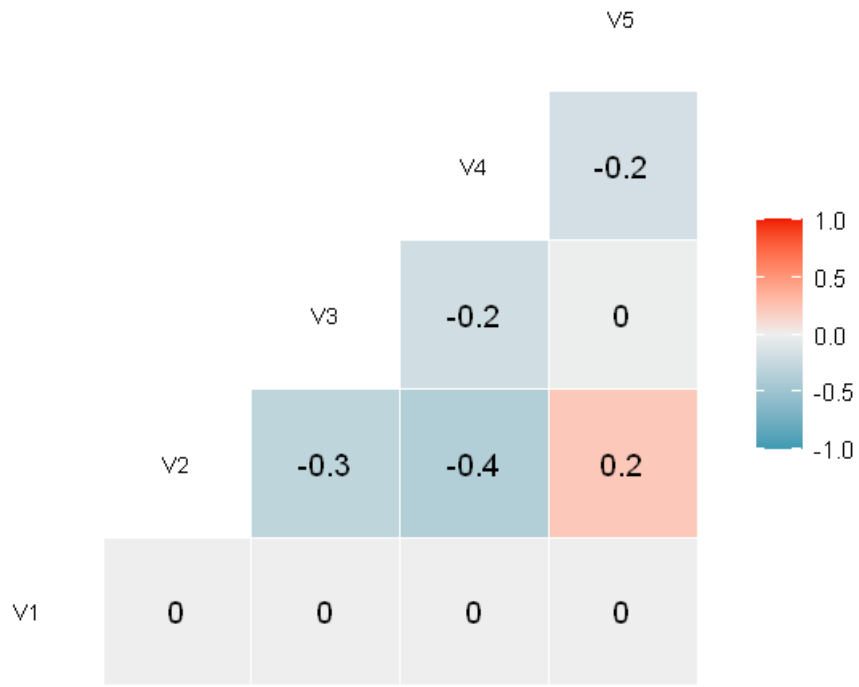
So we drop these variables from our model.

Hence, we re-define model2 based on removing all insignificant variables (including pctymle and others)

[41]:
```
model2 = lm(y ~ prbarr + prbconv + log10_density + pctmin80, data=data)
model = model2
# MAKE SURE TO RE-RUN THE ABOVE FUNCTIONS FOR DIAGNOSTICS PLOTS, SHAPIRO TEST,␣
↪BP TEST ETC.
```

We re-run the above tests, diagnostics and checks for this updated model2. Those result diagrams and diagnostics are what we observe above. Next we test for causality by establishing exogeneity in our updated model 2

[45]:
```
exodata = cbind(model$residuals, data$prbarr, data$prbconv, data$log10_density,␣
↪data$pctmin80)
ggcorr(exodata, size = 3, label=TRUE)
```



1.5.6 Causality Criteria

We notice that the residuals mean is 0. We also notice that the correlation between the residuals and the "significant" model2 variables is 0 for all variables (see last row in the above correlation matrix plot).

Hence, we can say that the exogeneity assumption is satisfied and hence, the effect of our model specification is causal

1.5.7 Model3: Maximalist approach model with backward elimination

[48]:
```
model3 = lm(y ~ . , data = data[, 4:26])
# summary(model3)

# step(model3, direction = "backward")
formula(model3)
```

```
y ~ prbarr + prbconv + prbpris + avgsen + polpc + density + taxpc +
    west + central + urban + pctmin80 + wcon + wtuc + wtrd +
    wfir + wser + wmfg + wfed + wsta + wloc + mix + pctymle
```

The stepsize variable selection procedure converges to the above model when we run the stepsize selection in with backward elimination.

For completeness, we also did a forward selection, and mixed approach (backward and forward). However, we decided to settle on the backward elimination model for parsimony and better fit.

1.5.8 Model Selection

We decided to select our balanced model **(model2 = y ~ prbarr + prbconv + log10_density + pctmin80)** with these variables because of a combination of reasons. 1. We found that this model had a number of variables that align well with our intuition of what typically may affect crime rate. Variables of criminal justice, along with some demographic variables are known in literature to influence crime rate. 2. We ran the model with labor market variables, including individual industry average wages. We found very little practical significance (beta values were very small) - so we determined labor market variables dont contribute much in influencing crime rate. Just to be sure, we conducted joint significance test for the wage variables, but the wald test showed us that the wage variables indeed dont contribute to the crime rate. So we modified model2 to use the above highlighted model. 3. Lastly, this model was very parsimonious and explained a fair amount of the variation for crime rate, so it was a good minimalist model, which fit well with "Occam's razor" principle.

1.6 6. Model Analysis

We notice that we have coefficients and std. errors from heteroskedastic robust methods (White-Huber)

We also have our regular std. errors from the summary command.

Note that the Breusch Pagan test shows absence of homoskedasticity (null hypothesis was rejected). However, if we look at the residual vs fitted plot, we can roughly conclude that the data is fairly homosekdastic.

The Shapiro test shows that errors are NOT normally distributed (at the 5% significance level). However, if we look at the QQ plot, we observe the data is fairly normal, except for a few datapoints with negative residuals.

We notice there are a few factors that are associated with the crime rate percentage (target is log10(crmrate)): 1. **Probability of Arrest (prbarr)** - this is fairly obvious. It has a negative coefficient, meaning that when probability of arrest goes up, crime rate percentage goes down - likely because criminals reconsider doing the crime, when they know they have a high chance of getting arrested. This variable is significant in both methods of coefficient calculation. 2. **Probability of Conviction (prbconv)** - This has a similar explanation to prbarr. 2. **Percentage of Minority (pctmin80)** - the model shows that as minorities increase in a county, so does crime. This variable is significant in both modes of coefficient calculation. What may be an interesting discussion is to find out whether the minorities are victims of crime, thus leading to higher crime rate OR are minorities the perpetrators of crime, thus leading to higher crime rate? 3. **Density (log10_density)** - as density goes up, crime rate goes up. This is consistent with our originally stated hypothesis - crowded regions have more chance for conflict and more chance of crime. 4. **Police per Capita (polpc)** (From model 1) - Although the polpc is not significant as a variable, it has a very large coefficient value, indicating this variable has practical significance! There is a temporal factor between crime rate and polpc (as observed in literature), which gives a delayed effect of police on crime rate. Hence, over a period of time, this practical significance may start showing some statistical significance.

Model fit metrics are looking good. Adjusted R squared is around 70% and AIC / BIC scores are negative (interesting to note).

1.7 7. Regression Table

[49]:
```
# finally plot the models and compare models with stargazer
se_model1 = sqrt(diag(vcovHC(model1)))
se_model2 = sqrt(diag(vcovHC(model2)))
se_model3 = sqrt(diag(vcovHC(model3)))

stargazer(model1, model2, model3, type = "text", omit.stat = "f",
          se = list(se_model1, se_model2, se_model3),
          star.cutoffs = c(0.05, 0.01, 0.001))
```

```
==========================================================================
                                  Dependent variable:
                     ---------------------------------------------
                                          y
                           (1)             (2)             (3)
--------------------------------------------------------------------------
log10_density           0.293***        0.365***
                         (0.070)         (0.060)
```

prbarr	-0.807***	-0.851***	-0.797***
	(0.169)	(0.174)	(0.165)
prbconv	-0.572***	-0.571***	-0.592***
	(0.157)	(0.134)	(0.175)
prbpris			-0.120
			(0.194)
avgsen			-0.006
			(0.006)
pctmin80	0.571***	0.581***	0.391**
	(0.077)	(0.077)	(0.138)
wcon			0.0002
			(0.0003)
wtuc			0.0001
			(0.0003)
wtrd			0.0003
			(0.001)
wfir			-0.0004
			(0.001)
wser			-0.001
			(0.001)
wmfg			-0.0001
			(0.0002)
wfed			0.001*
			(0.0005)
wsta			-0.0005
			(0.0004)
wloc			0.001
			(0.001)
mix			-0.159
			(0.228)
pctymle			1.163*
			(0.563)

```
polpc                    102.310                          106.402
                        (55.229)                         (65.110)

density                                                   0.048*
                                                         (0.023)

taxpc                                                     0.001
                                                         (0.003)

west                                                     -0.051
                                                         (0.054)

central                                                  -0.048
                                                         (0.041)

urban                                                    -0.047
                                                         (0.097)

Constant                -1.487***       -1.315***        -1.747***
                         (0.121)         (0.092)          (0.330)

-----------------------------------------------------------------------
Observations               88              88               88
R2                        0.815           0.752            0.839
Adjusted R2               0.803           0.741            0.785
Residual Std. Error 0.098 (df = 82) 0.112 (df = 83) 0.102 (df = 65)
=======================================================================
Note:                                  *p<0.05; **p<0.01; ***p<0.001
```

We compare the results between the 3 models. We notice that while model2 (updated balanced model) has the lowest Adjusted R-squared value, it in fact is achieving the best performance with maximum parsimony.

All variables in model2 are highly statisically significant.

Model1 has polpc as a variable, that is not significant, but it improves the adjusted R squared because of high practical significance as well as the chance that polpc may influence crime rate over a period of time.

Model3 has a high adjusted R squared, however, this is expected because the number of variables is high, compared to the other 2 models - so this is expected.

1.8 8. Assumptions of Multiple Linear Regression

1. Parameters are linear
2. No perfect multicollinearity
3. IID / Random sample of observations in dataset
4. Zero conditional mean, ie, E[u | X] = 0 (or weaker assumption of exogeneity, E(u) = 0 and Cov(X, u) = 0)
5. Homoskedastic errors, Var(u | X) = const

6. Normality of errors ~ N(0, sigmaˆ2)

MLR 1-4 implies unbiased OLS estimators (beta coefficients) MLR 1-5 (Gauss markov) implies OLS estimators are BLUE MLR 1-6 (Classical LM assumptions) implies OLS estimators, ie, beta_j ~ N(beta_j, var(beta_j)) - var(beta_j) given by white standard errors

1.8.1 MLR assumptions for Modified Model 2

1. **MLR 1 - linear in parameters** is satisfied for all models based on our EDA between our target variable and explanatory variables.
2. There is **no PERFECT multicollinearity** in the data as we have removed variables that could have potentially shown high correlation. Also VIF score for all explanatory variables < 4. **Hence MLR 2 is also satisfied.**
3. The data is assumed to be collected as IID (this can be assumed from the study). The authors of the study that produced this data claimed to meet this assumption (sampling method). The fact that our data consists of a slice of this data from a study may suggest that we cannot say this assumption is met. For example, it is possible that the residuals from a single year (which is the case for our dataset) might not be IID. If there is some change in these data over time, having temporal data would allow us to investigate this possibility. The geographic proximity of these counties in North Carolina further complicates our ability to assert that our residuals are IID. The authors of the original study employed advanced methods to correct for this problem. As these methods are beyond the scope of this class, we can only note that this assumption is not met. **Hence, MLR 3 criteria is assumed to be satisfied.**
4. Looking at the residuals vs fitted values plot, **MLR 4 (Zero Conditional mean) is satisfied.** The zero conditional mean line does slope downwards on extremities, however, this could be due to lack of sufficient data points.
5. The Breusch Pagan test shows us that the variance of errors is barely heteroskedastic (at the 95% signifcance level). This is also confirmed from the std. residuals plot. **This shows us that MLR 5 (homoskedasticity) is approx. satisfied.**
6. The QQ norm plot and residual histogram plot for errors shows that errors do follow an approximate normal distribution. We also calculate the mean of errors to show that we have zero residual mean. However, there are some clear outliers represented in the data (skewed towards the lower end). **This shows that MLR 6 (normality of errors) is also satisfied.**
7. Because we have 88 data points (> 30), we can rely on the **law of asymptotics (CLT)** to say that even if MLR 6 is not completely satisfied, the coefficients are normally distributed with unbiased true population parameter value as mean and variance given by White-Huber standard errors, ie, betaj_hat ~ N(beta_j, var(beta_j)).

1.8.2 Outliers

1. We notice that there are a few outliers in our data.
2. These also show up in the QQ plot as well as the Cooks distance plot (having cook's dist close to 1). However, we will include these in our model because its not much larger than 1.
3. **Note that we removed some rows which had an extremely large cook's distance and was causing problems in the model.**

1.9 9. Omitted Variables

We pick 3 county wide variables as omitted variables: 1. Drug_Pct - the percentage of people in the county using drugs 2. Unemployment_Pct - the percentage of unemployed people 3. Education_Years - the average number of years of education in the county 4. Weather - the degrees celcius temperature in the county 5. Household_Wealth - net worth of the household.

We expect Drug_Pct to be positively correlated with crime rate. We expect Unemployment_Pct to be positively correlated with crime rate. We expect Education_Years to be negatively correlated with crime rate. We expect Weather to be positively correlated with crime rate. We expect Household_Wealth to be negatively correlated with crime rate.

For the omitted variable bias, we need to consider our modified beta values (alphaj):

alphaj = betaj + beta_omitted x delta

where alphaj is off from the population betaj by a factor of **beta_omitted x delta**

Let us say our jth variable is prbarr

In the case of (say) Education_Years, 1. beta_omitted is negative (Education_Years should be negatively correlated with crime rate) 2. delta is negative (prbarr should be negatively correlated with Education_Years)

Hence, beta_omitted x delta is positive. Coefficient for prbarr is negative and we add a positive value (beta_omitted x delta) to it. So we have drawn the new coefficient (alphaj) for prbarr closer to 0.

If Education_Years was included, betaj would be less negative.

Similarly, for Drug_Pct variable, the (beta_omitted x delta) bias is overall positive and it will make the new co-efficient less negative like Education.

For Unemployment_Pct variable, the (beta_omitted x delta) bias is overall positive and it will make the new co-efficient less negative like Education.

For Weather variable, the (beta_omitted x delta) bias is overall positive and it will make the new co-efficient less negative like Education.

For Household_Wealth variable, the (beta_omitted x delta) bias is overall positive and it will make the new co-efficient less negative like Education.

1.10 10. Conclusion

Crime in North Carolina clearly has a complex set of root causes and effects but here we can draw a few conclusions that could help potentially drive several policy initiatives to reduce the crime rate in many counties.

The first set of policy initiatives we would suggest would be around strengthening certain aspects of the criminal justice system. The second set of policy initiatives may include influencing certain demographic issues of the counties.

We know that an increase in the probability of arrest and conviction will result in lower crime rates. So ensure that perpetrators are identified correctly and arrested; The prosecutor has better evidence that can lead to higher convictions and thereby lower crime rates. For this, we will need to invest in camera technology and hire staff and police personnel to monitor and follow up on criminal behavior captured on cameras.

Population demographics are clearly at play here. We found a relationship between a dense population with higher percentages of minorities correlated with a higher crime rate. While we can take subtle measures like making counties less congested (sparse housing), a more practical suggestion is to begin to help Minority community restore faith in the Justice system. Based on information from EJI (Equal Justice Initiative), In 1972 there were 200K people incarcerated. In

2017 there were 2.2 Million. There has been a 750% increase of women in jails in the since 1980. Majority of the convictions are minorities; Children in Prisons, Wrongful convictions, excessive punishment and Prison conditions, all result in the community loss of faith in the justice system. Self improvement and support programs are needed to restore the trust and get minority communities involved in fighting and preventing crime. These may divert potential criminal activity into more productive activities with life long benefits, avoiding the stigma of a criminal record for many.

While wages were partially significant in our model, further analysis is needed to establish if wealth and economic factors actually affect our population model for crime. Building out this analysis with more data around wealth, such as the level of assets in a county, how many families own homes and the level and access to higher education would make such as study more fruitful.

Next steps for follow-on research. Three areas in which we need more data and analysis for each county are: Economic opportunity, education, household wealth, and prevalence of drugs. Temporal data around how the crime rate changes with these variables every year is also important to have to capture temporal effects of these variables, and other known variables like police per capita. These factors will give more information on the impact of these factors and then we will be able to formulate policy to reduce these as contributors to the crime in these communities.

[ ]: