Statistics for Crime Science

Lab 3: Reducing Crime - Louis Li, Chris Sexton, Carver Sorensen

November 25th 2019

Introduction

Crime Rate in North Carolina is a potential topic for a political campaign. Candidates need to have an opinion on what impacts the rate of crime across the state and need to be able to discuss the policy decisions that could lead to reducing crime rate in North Carolina. This report seeks to investigate the contributing factors to regional crime rate and to make political and policy recommendations to reduce these rates.

The team looks into data collected from a research project undertaken by the University of Georgia and West Virginia University in 1994. In particular it looks at the impact of population density and policing policy in the state and whether adjusting resources or policy in these areas can have a positive impact on reducing crime.

We will address this issue by using exploratory data analysis (EDA) techniques and linear regression modelling. Through this exercises we will build a number of models that can be used to form policy decisions and talking points.

The data is taken from a study by Cornwell and Trumball, researchers from the University of Georgia and West Virginia University (C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data," Review of Economics and Statistics 76, 360-366.). Each row of data represents statistics for each county in North Carolina.

We will present three linear regression models in this report. Model 1 will be our initial base model looking at only one or two variables. The second model, Model 2, will look at variables that we think have the most important impact on crime rate in line with our investigation on policy and population density. The third model will be more broad ranging and consider all, or most of the available variables available to us.

```
## [1] "Number of observations 97"
```

```
## [1] "Number of variables 25"
```

The data is taken from the year 1987 and contains 97 observations of 25 variables that could be related to the rate of crime in North Carolina.

This report consists of an EDA section, which describes the sample data collected and looks to provide some evidence of variable selection and transformation. After this, the report shows and discusses the results of linear regression models that could be used to find relationship between the chosen variables and crime rate. Finally, the report concludes with recomendations from the EDA and modelling.

Explanatory Data Analysis (EDA)

The first step in the analysis is to view the types and structure of the data, initially getting a simple overview:

```
## 'data.frame':    97 obs. of  25 variables:
##  $ county  : int  1 3 5 7 9 11 13 15 17 19 ...
##  $ year    : int  87 87 87 87 87 87 87 87 87 87 ...
##  $ crmrte  : num  0.0356 0.0153 0.013 0.0268 0.0106 ...
##  $ prbarr  : num  0.298 0.132 0.444 0.365 0.518 ...
##  $ prbconv : Factor w/ 92 levels "","`","0.068376102",..: 63 89 13 62 52 3 59 78 42 86 ...
##  $ prbpris : num  0.436 0.45 0.6 0.435 0.443 ...
##  $ avgsen  : num  6.71 6.35 6.76 7.14 8.22 ...
##  $ polpc   : num  0.001828 0.000746 0.001234 0.00153 0.00086 ...
##  $ density : num  2.423 1.046 0.413 0.492 0.547 ...
##  $ taxpc   : num  31 26.9 34.8 42.9 28.1 ...
##  $ west    : int  0 0 1 0 1 1 0 0 0 0 ...
##  $ central : int  1 1 0 1 0 0 0 0 0 0 ...
##  $ urban   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ pctmin80: num  20.22 7.92 3.16 47.92 1.8 ...
##  $ wcon    : num  281 255 227 375 292 ...
##  $ wtuc    : num  409 376 372 398 377 ...
##  $ wtrd    : num  221 196 229 191 207 ...
##  $ wfir    : num  453 259 306 281 289 ...
##  $ wser    : num  274 192 210 257 215 ...
##  $ wmfg    : num  335 300 238 282 291 ...
##  $ wfed    : num  478 410 359 412 377 ...
##  $ wsta    : num  292 363 332 328 367 ...
##  $ wloc    : num  312 301 281 299 343 ...
##  $ mix     : num  0.0802 0.0302 0.4651 0.2736 0.0601 ...
##  $ pctymle : num  0.0779 0.0826 0.0721 0.0735 0.0707 ...
```

Most of the data are type numeric, which is useful for our analysis. The prbconv (probability of conviction) is a Factor data type with 92 levels. As probabilities should be float values (ideally from 0 to 1), the Factor type must be changed to numeric.

The west, central and urban district variables are integers and look to be only 0 and 1, which seems correct. We will need to make sure this is true for all values of these variables.

The year appears to be constant at 87; if this is the case, then we can ignore this variable.

We perform a summary of the data to check in more detail.

```
##      county           year        crmrte             prbarr
##  Min.   :  1.0   Min.   :87   Min.   :0.005533   Min.   :0.09277
##  1st Qu.: 52.0   1st Qu.:87   1st Qu.:0.020927   1st Qu.:0.20568
##  Median :105.0   Median :87   Median :0.029986   Median :0.27095
##  Mean   :101.6   Mean   :87   Mean   :0.033400   Mean   :0.29492
##  3rd Qu.:152.0   3rd Qu.:87   3rd Qu.:0.039642   3rd Qu.:0.34438
##  Max.   :197.0   Max.   :87   Max.   :0.098966   Max.   :1.09091
##  NA's   :6       NA's   :6    NA's   :6          NA's   :6
##         prbconv      prbpris           avgsen           polpc
##             : 5   Min.   :0.1500   Min.   : 5.380   Min.   :0.000746
##  0.588859022: 2   1st Qu.:0.3648   1st Qu.: 7.340   1st Qu.:0.001231
##  `          : 1   Median :0.4234   Median : 9.100   Median :0.001485
##  0.068376102: 1   Mean   :0.4108   Mean   : 9.647   Mean   :0.001702
##  0.140350997: 1   3rd Qu.:0.4568   3rd Qu.:11.420   3rd Qu.:0.001877
##  0.154451996: 1   Max.   :0.6000   Max.   :20.700   Max.   :0.009054
##  (Other)    :86   NA's   :6        NA's   :6        NA's   :6
##     density          taxpc             west           central
##  Min.   :0.00002   Min.   : 25.69   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.54741   1st Qu.: 30.66   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.96226   Median : 34.87   Median :0.0000   Median :0.0000
##  Mean   :1.42884   Mean   : 38.06   Mean   :0.2527   Mean   :0.3736
##  3rd Qu.:1.56824   3rd Qu.: 40.95   3rd Qu.:0.5000   3rd Qu.:1.0000
##  Max.   :8.82765   Max.   :119.76   Max.   :1.0000   Max.   :1.0000
##  NA's   :6         NA's   :6        NA's   :6        NA's   :6
##      urban            pctmin80           wcon            wtuc
##  Min.   :0.00000   Min.   : 1.284   Min.   :193.6   Min.   :187.6
##  1st Qu.:0.00000   1st Qu.: 9.845   1st Qu.:250.8   1st Qu.:374.6
##  Median :0.00000   Median :24.312   Median :281.4   Median :406.5
##  Mean   :0.08791   Mean   :25.495   Mean   :285.4   Mean   :411.7
##  3rd Qu.:0.00000   3rd Qu.:38.142   3rd Qu.:314.8   3rd Qu.:443.4
##  Max.   :1.00000   Max.   :64.348   Max.   :436.8   Max.   :613.2
##  NA's   :6         NA's   :6        NA's   :6       NA's   :6
##       wtrd            wfir            wser             wmfg
##  Min.   :154.2   Min.   :170.9   Min.   : 133.0   Min.   :157.4
##  1st Qu.:190.9   1st Qu.:286.5   1st Qu.: 229.7   1st Qu.:288.9
##  Median :203.0   Median :317.3   Median : 253.2   Median :320.2
##  Mean   :211.6   Mean   :322.1   Mean   : 275.6   Mean   :335.6
##  3rd Qu.:225.1   3rd Qu.:345.4   3rd Qu.: 280.5   3rd Qu.:359.6
##  Max.   :354.7   Max.   :509.5   Max.   :2177.1   Max.   :646.9
##  NA's   :6       NA's   :6       NA's   :6        NA's   :6
##       wfed            wsta            wloc            mix
##  Min.   :326.1   Min.   :258.3   Min.   :239.2   Min.   :0.01961
##  1st Qu.:400.2   1st Qu.:329.3   1st Qu.:297.3   1st Qu.:0.08074
##  Median :449.8   Median :357.7   Median :308.1   Median :0.10186
##  Mean   :442.9   Mean   :357.5   Mean   :312.7   Mean   :0.12884
##  3rd Qu.:478.0   3rd Qu.:382.6   3rd Qu.:329.2   3rd Qu.:0.15175
##  Max.   :598.0   Max.   :499.6   Max.   :388.1   Max.   :0.46512
##  NA's   :6       NA's   :6       NA's   :6       NA's   :6
##     pctymle
##  Min.   :0.06216
##  1st Qu.:0.07443
##  Median :0.07771
##  Mean   :0.08396
##  3rd Qu.:0.08350
##  Max.   :0.24871
##  NA's   :6
```

Cleaning

From the summary we can see some obvious errors that need fixing.

- There are 6 NAs which means some data are missing.
- prbconv has blank values and a non-numeric value " ‘ ".
- The data type of prbconv is non-numeric due to data errors.
- prbarr has a max value greater than 1. A chance of being arrested greater than one does not make sense intuitively, therefore we will need to examine more closely how this variable is defined.

To fix the above we first remove the NAs; this also removes the rows with non-numeric prbconv values.

Check if NAs are consistent across all rows:

```
##    county year crmrte prbarr prbconv prbpris avgsen polpc density taxpc
## 92     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 93     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 94     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 95     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 96     NA   NA     NA     NA              NA     NA    NA      NA    NA
## 97     NA   NA     NA     NA       `      NA     NA    NA      NA    NA
##    west central urban pctmin80 wcon wtuc wtrd wfir wser wmfg wfed wsta
## 92   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 93   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 94   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 95   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 96   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
## 97   NA      NA    NA       NA   NA   NA   NA   NA   NA   NA   NA   NA
##    wloc mix pctymle
## 92   NA  NA      NA
## 93   NA  NA      NA
## 94   NA  NA      NA
## 95   NA  NA      NA
## 96   NA  NA      NA
## 97   NA  NA      NA
```

We remove NAs and turn prbconv into numeric values, which is consistent with other probability variables. The histogram in Figure 1 shows the distribution of the amended prbconv variable.

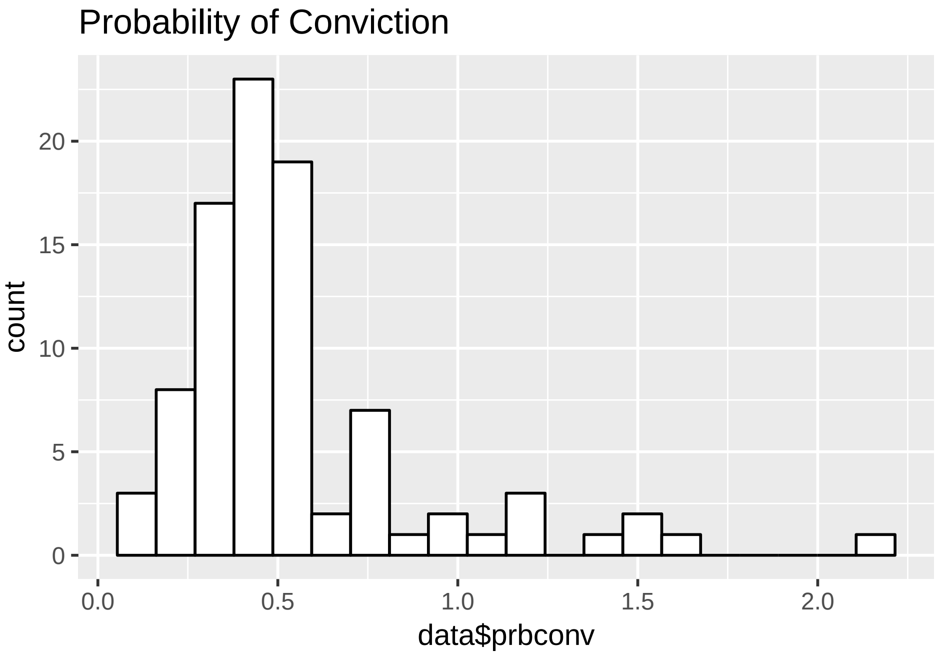


Figure 1: Histogram of Probability of Conviction

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

The summary data shows that there are prbconv values above 1, which could be an anomoly. We check how many prbconv values are above 1 to see how widespread the issue is.

```
##    county year    crmrte   prbarr prbconv  prbpris avgsen      polpc
## 1       3   87 0.0152532 0.132029 1.48148 0.450000   6.35 0.00074588
## 2      19   87 0.0221567 0.162860 1.22561 0.333333  10.34 0.00202425
## 3      99   87 0.0171865 0.153846 1.23438 0.556962  14.75 0.00185912
## 4     115   87 0.0055332 1.090910 1.50000 0.500000  20.70 0.00905433
## 5     127   87 0.0291496 0.179616 1.35814 0.335616  15.99 0.00158289
## 6     137   87 0.0126662 0.207143 1.06897 0.322581   6.18 0.00081426
## 7     149   87 0.0164987 0.271967 1.01538 0.227273  14.62 0.00151871
## 8     185   87 0.0108703 0.195266 2.12121 0.442857   5.38 0.00122210
## 9     195   87 0.0313973 0.201397 1.67052 0.470588  13.02 0.00445923
## 10    197   87 0.0141928 0.207595 1.18293 0.360825  12.23 0.00118573
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 1  1.0463320 26.89208    0       1     0  7.91632 255.1020 376.2542
## 2  0.5767442 61.15251    0       0     0 24.31170 260.1381 613.2261
## 3  0.5478615 39.57348    1       0     0 14.28460 259.7841 417.2099
## 4  0.3858093 28.19310    1       0     0  1.28365 204.2206 503.2351
## 5  1.3388889 32.02376    0       0     0 34.27990 290.9091 426.3901
## 6  0.3167155 44.29367    0       0     0 33.04480 299.4956 356.1254
## 7  0.6092437 29.03402    1       0     0 10.00460 223.6136 437.0629
## 8  0.3887588 40.82454    0       1     0 64.34820 226.8245 331.5650
## 9  1.7459893 53.66693    0       0     0 37.43110 315.1641 377.9356
## 10 0.8898810 25.95258    1       0     0  5.46081 314.1660 341.8803
##        wtrd     wfir      wser   wmfg   wfed   wsta   wloc        mix
## 1  196.0101 258.5650  192.3077 300.38 409.83 362.96 301.47 0.03022670
## 2  191.2452 290.5141  266.0934 567.06 403.15 258.33 299.44 0.05334728
## 3  168.2692 301.5734  247.6291 258.99 442.76 387.02 291.44 0.01960784
## 4  217.4908 342.4658  245.2061 448.42 442.20 340.39 386.12 0.10000000
## 5  257.6008 441.1413  305.7612 329.87 508.61 380.30 329.71 0.06305506
## 6  170.8711 170.9402  250.8361 192.96 360.84 283.90 321.73 0.06870229
## 7  188.7683 353.2182  210.4415 289.43 421.34 342.92 301.23 0.11682243
## 8  167.3726 264.4231 2177.0681 247.72 381.33 367.25 300.13 0.04968944
## 9  246.0614 411.4330  296.8684 392.27 480.79 303.11 337.28 0.15612382
## 10 182.8020 348.1432  212.8205 322.92 391.72 385.65 306.85 0.06756757
##       pctymle
## 1  0.08260694
## 2  0.07713232
## 3  0.12894706
## 4  0.07253495
## 5  0.07400288
## 6  0.07098370
## 7  0.06215772
## 8  0.07008217
## 9  0.07945071
## 10 0.07419893
```

The team identifies there are 10 instances where the value for prbconv is greater than 1. It is possible for an individual to be arrested once and then convicted for multiple offenses/counts. This could realistically cause the ratio of convictions to arrests to be greater than 1. Therefore, the team decides not to change any of the values for prbconv at this time. Outliers are not initially seen from the histogram, but this variable will be explored further in the model section.

As noted in the summary of variables, there is at least one prbarr that is greater than 1. The variable prbarr is defined as the ratio of arrests over offenses. Although it is not intuitive for a probability of arrests to be greater than one, it is possible for multiple arrestes to be made for the same offensive, and therefore the ratio of arrests against offenses can in theory be greater than 1. The histogram in Fugure 2 shows the distribution for prbarr:

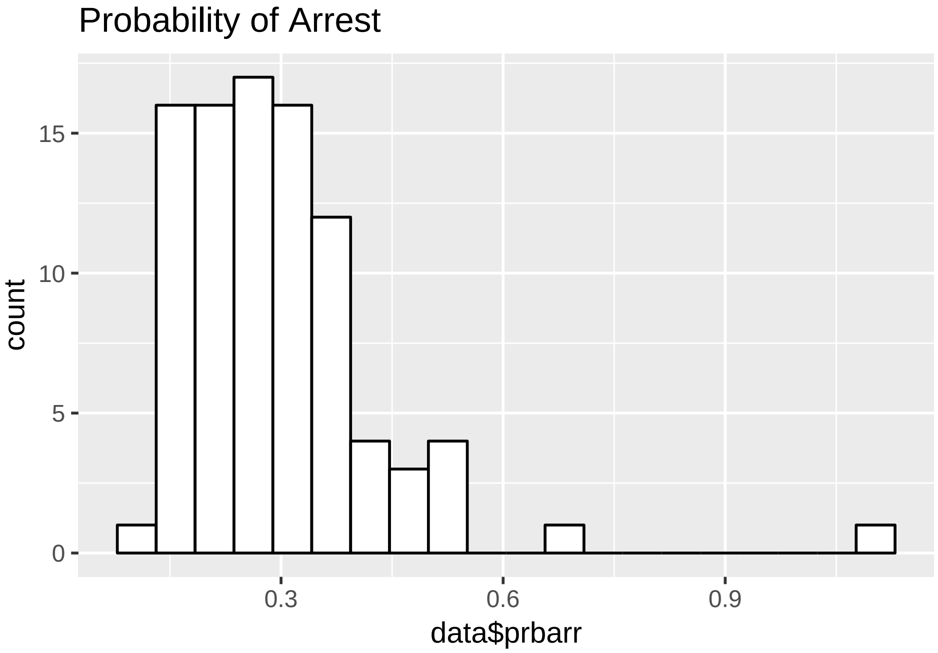


Figure 2: Histogram of Probability of Arrest

Now we check which instances of prbarr have values above 1.

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

```
## [1] "Number of standard deviations county 115's prbarr is from the mean:  5.81"
```

```
## [1] "Mean of other counties:  0.29"
```

There is only one row with a prbarr greater than 1, county 115, with a probability of arrest at 1.09091. The mean value of all other counties is 0.29, which reinforces the argument that county 115's prbarr value could be an error.

We have options in how we deal with this potentially erroneous value:

- Assume it is correct and leave it as is. It seems unlikely that it is possible to get a probability of arrests for a county greater than 100%, but possible based on the definition.
- Remove the row. The rest of the records appear to be legitimate, and we do not want to delete values unnecessarily
- Use a prediction method such as KNN or cosine similarity to predict the value. This is a legitimate option but outside the technical scope of this analysis.
- Use a geographical value from a neighboring county. It seems that the county IDs represent FIPS codes, and thus it is possible to calculate a prbarr value from neighboring counties. This method is rejected on the basis that we cannot assume arrest probabilities are similar from one neighboring county to the next.

- Use the highest possible value. We cannot know that this variable was designed to be a 1, and in fact that would be highly unlikely.
- Take the mean, median or mode of all counties and use this result

Initially the team decided to use the last method and adjust the variable based on the assumtion it was an error. However on further examination, the team concluded that it is possible for the prbarr to be greater than 1 and therefore it will be left in for now. However this observation will be re-examined in model evaluation with specific attention to measuring Cook's distance for outlier detection.

```
## [1] "Number of rows:  91"
```

```
## [1] "Number of columns:  25"
```

Our sample size is reduced to 91 observations having removed the NA rows. We now need to check for duplicate rows.

```
##    county year    crmrte   prbarr  prbconv  prbpris avgsen      polpc
## 89    193   87 0.0235277 0.266055 0.588859 0.423423   5.86 0.00117887
##      density    taxpc west central urban pctmin80     wcon     wtuc
## 89 0.8138298 28.51783    1       0     0  5.93109 285.8289 480.1948
##        wtrd     wfir     wser   wmfg   wfed   wsta   wloc       mix
## 89 268.3836 365.0196 295.9352 295.63 468.26 337.88 348.74 0.1105016
##       pctymle
## 89 0.07819394
```

There is one duplicate row which we will remove.

```
## [1] "Number of rows:  90"
```

```
## [1] "Number of columns:  25"
```

The year variable has a single repeated value, "87", throughout the dataset. This variable can be omitted from the analysis, as it does not provide any useful information for the modelling.

Outliers

We now check for outliers in the general data set. We are mainly interested if there are any obvious outlier errors in the data and to be aware of potential issues. A histogram gives some initial indication. We ignore the variables that are binary in nature (county, west, central and urban). We will look more closely at outliers when we produce models.

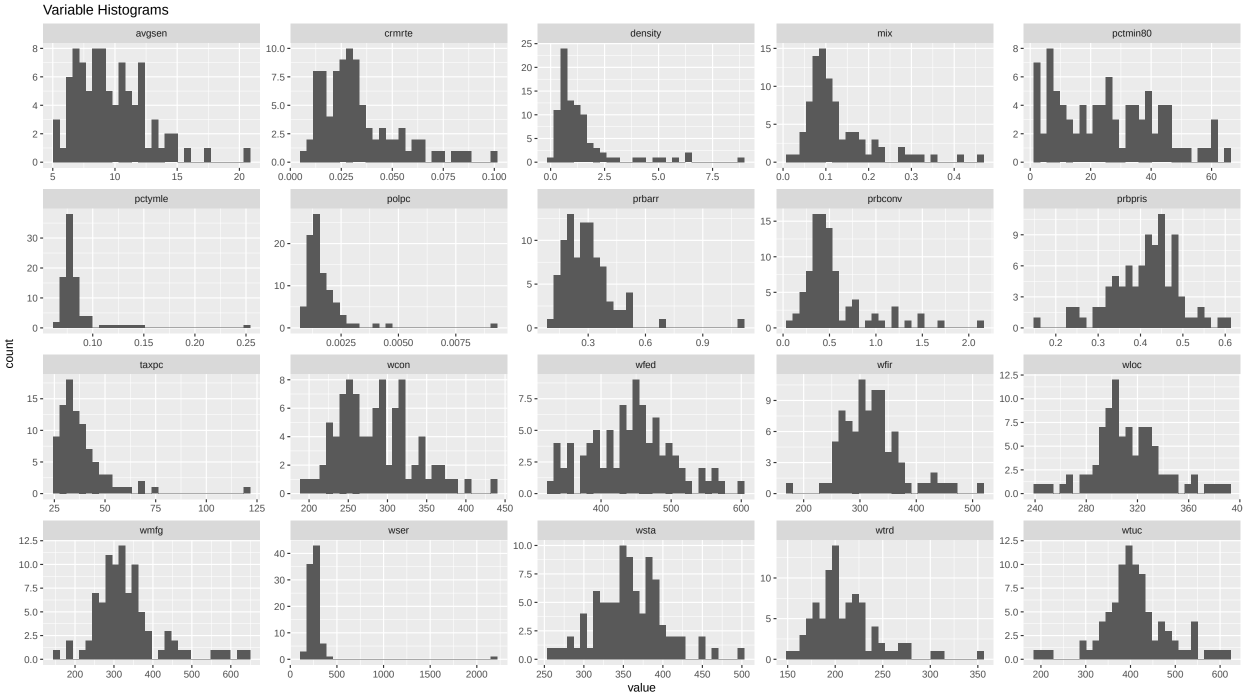


Figure 3: Histograms of Variables

The histograms in Figure 3 show that we have a variation of distributions; we are primarily concerned with data related to policy and density population as related to crime rate. We include histograms for other variables to give an overall understanding of the data and potential understanding for later models, especially Model 3.

- avgsen: Average Sentence. The data are not normally distributed and has a slight positive skew. Using the log of avgsen may make sense, considering average sentence as a percentage change rather than an actual number could be useful both for interpretation and performance of modelling.
- crmrte: The number of crimes committed per person. These are also slightly positively skewed. A log of this variable may also be useful for performance purposes however it is slightly harder, although not impossible, to interpret the results. The log of crime rate would represent the percentage change in the rate of crime. Overall this is still an intuitive variable.
- pctmin80: Percent minority in 1980. This is a variable could be part of our population study as it considers the percentage of minorities in a county, however the data range of 1980 seems too far away from our results taken from 1987.
- pctymle: Percentage of young males in a county. This looks to have an outlier, otherwise a relatively normal distribution.
- polpc: The police per capita. This also looks to have an outlier
- prbpris: 'probability' prison sentence. This variable is normally distributed.
- pbarr: 'probability' of arrest. This variable has an outlier which is pulling the distribution positively. Removing this outlier would likely lead to a normal distribution.
- prbconv: 'probability' of conviction is positively skewed with a number of observations above 1 as previously discussed. Removing the values greater than 1 would give a more normal distribution but we leave them in for now.
- wser- it should be noted that wser (weekly wage, service industry) has one very obvious outlier which we will examine further in the report.

We also take a look at the skewness scores, which will also help identify outliers.

Skewness Scores

	Skewness
crmrte	1.2817489
prbarr	2.5252960
prbconv	2.0395060
prbpris	-0.4525402
avgsen	1.0011634
polpc	4.9834880
density	2.6435077
taxpc	3.2905745
pctmin80	0.3656617
wcon	0.6068022
wtuc	0.0681977
wtrd	1.4612066
wfir	0.8206315
wser	8.6991816
wmfg	1.4225317
wfed	0.1322376
wsta	0.3623683
wloc	0.2951381
mix	1.9165705
pctymle	4.5606907

The polpc, wser and pctymle variables have high magnitudes for skewness, with taxpc as the next highest. This indicates that outliers may exist within these variables. We check further with a boxplot diagram.

Selected Outliers Boxplot

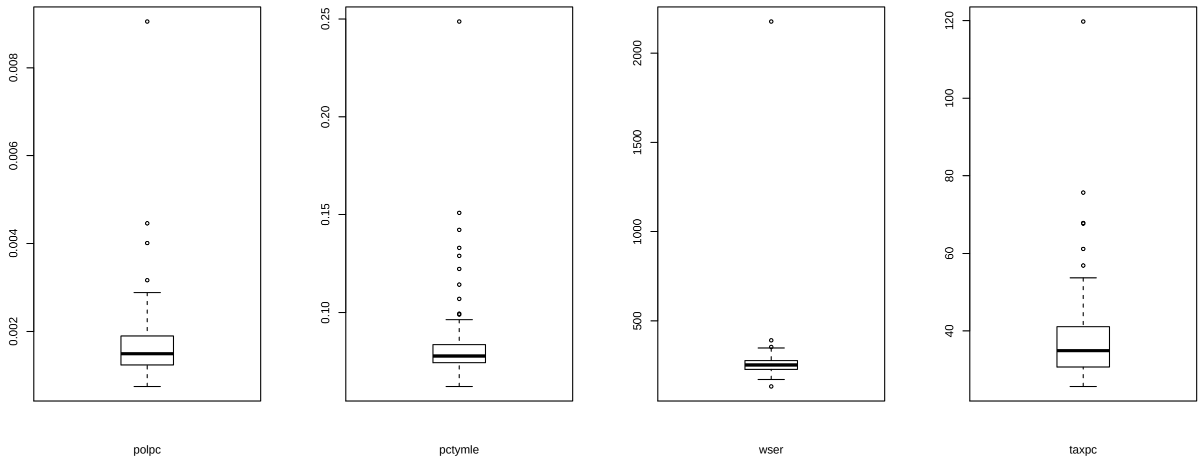


Figure 4: Outliers

All of these outliers may need to be addressed in the models if they are used, depending on their influence and leverage.

Furthermore, we can look at regional variables to make sure there are no erroneous records.

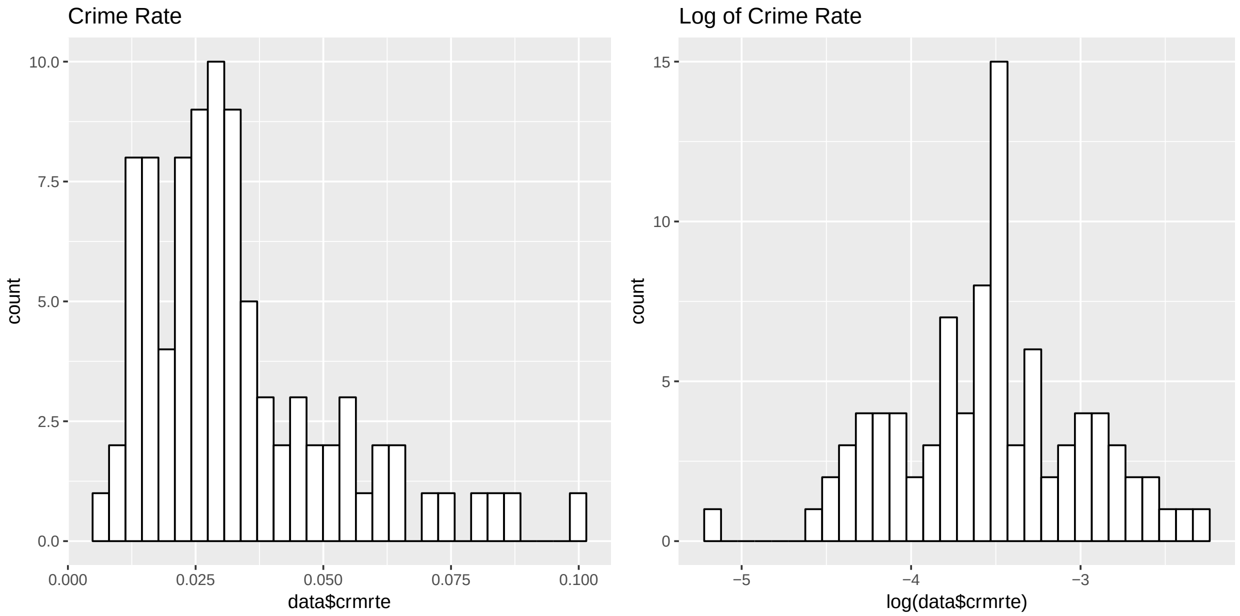
We see a few dummy variables - west, central and urban. West and Central are indicators of region in N.C and since there are other regions besides West and Central, we do not have a dummy variable trap. Similarly for Urban, there is no dummy variable trap.

```
##   county    crmrte   prbarr prbconv  prbpris avgsen      polpc  density
## 1     71 0.0544061 0.243119 0.22959 0.379175  11.29 0.00207028 4.834734
##      taxpc west central urban pctmin80     wcon     wtuc     wtrd     wfir
## 1 31.53658    1       1     0   13.315 291.4508 595.3719 240.3673 348.0254
##       wser   wmfg   wfed   wsta   wloc       mix    pctymle
## 1 295.2301 358.95 509.43 359.11 339.58 0.1018608 0.07939028
```

The team observes there is one county that is considered to be both central and west. According to our research county #71 is Gaston, part of central North Carolina. We decided not to change the value for this record.

Distribution and Correlation

Given that we want to address crime rate, there is a variable that directly addresses this - we will initially look to use crime rate our target variable. First we examine more closely the distribution of this variable.


Crime Rate

Log of Crime Rate

Figure 5: Crime Rate Distribution

As previously noted crmrte is positively skewed and does not have a normal distribution. Strictly speaking, the central limit theorem allows us to use the unmodified crmrte variable; however, we may wish to consider using the log of crmrte as our target variable, as noted by Wooldridge in Introductory Economics (2015) chapter 6 summary - “Models using the log(y) as the dependent variable will often more closely satisfy the classical linear model assumptions.” For example the model has a better chance of being linear, homoskedascity is more likely to hold and normality is often more plausible."

Having chosen the target varicable, we look discuess independent variables, we ignore some of these based on our understanding of their usefulness:

– polpc: We ignore the police per capita, as we think that the amount of police in an area is determined by the crime rate and not the other way around, therefore polpc is a dependent variable on crmrte. We cannot be 100% sure if this is true, but intuitively we think this is likely and therefore it is not used in our modelling.

– mix: This variable relates to the type of crime committed, either face to face or other. We think it would be difficult (although not impossible) to direct policy towards certain types of crime. Additionally, this variable does not tell us much information. The bucket “other” is wide ranging, so we cannot direct any specific policy to address this. We will omit mix from our study.

– pctmin80: This variable describes the percentage of minorities in 1980, seven years prior to this sample. We do not know if this percentage is still relevant in 1987, but regardless, it does not define what minority is. Therefore, it would be difficult to direct policy towards this variable.

– urban, central, west: These variables are binary in nature, so the scatterplots would not reveal meaningful information. We will explore the boxplots of the log(crmrte) for each of these binary features.

The following three boxplot charts in Figure 6 describe the relationship of crime rate on the different location variables: urban, west and central. Note that the difference in mean crime rate for counties described as urban (factor = 1) are higher than non-urban (factor = 0). A similar situation, although to a lesser extent, occurs for western counties, and this also holds true to the central variable to an even lesser extent. The team will examine these variables in Model 2.

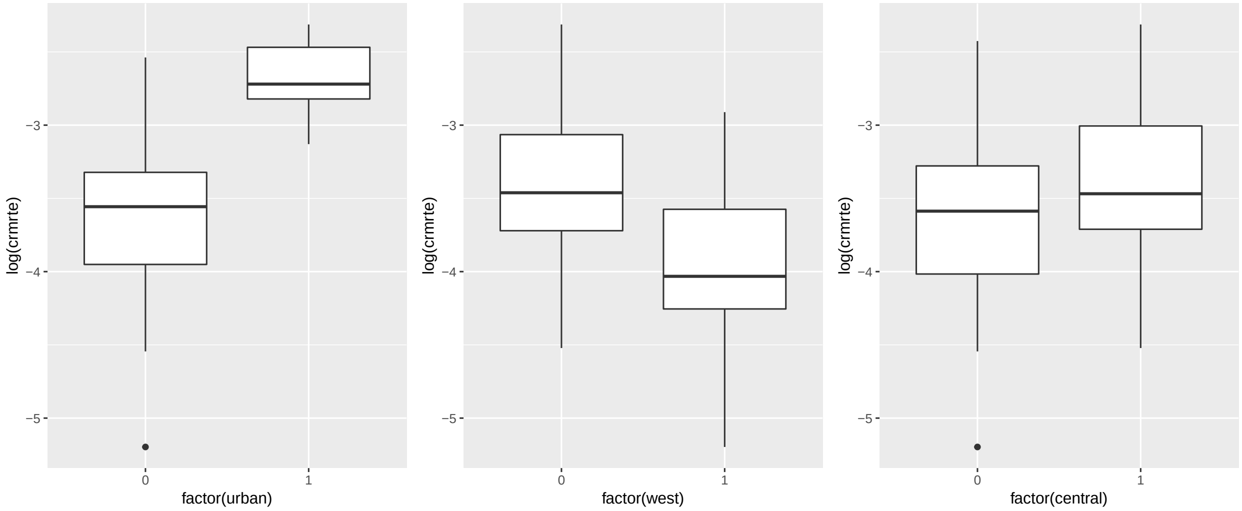


Figure 6: Crime Rate by Location

Given that we have a target variable, we can examine more closely the effect on the previously viewed outliers on crime rate. The following charts in Figure 7 show the Residual vs Leverage plots, specifically Cook’s distance, which tells us more concretely if the outlier observations are likely to affect a model.

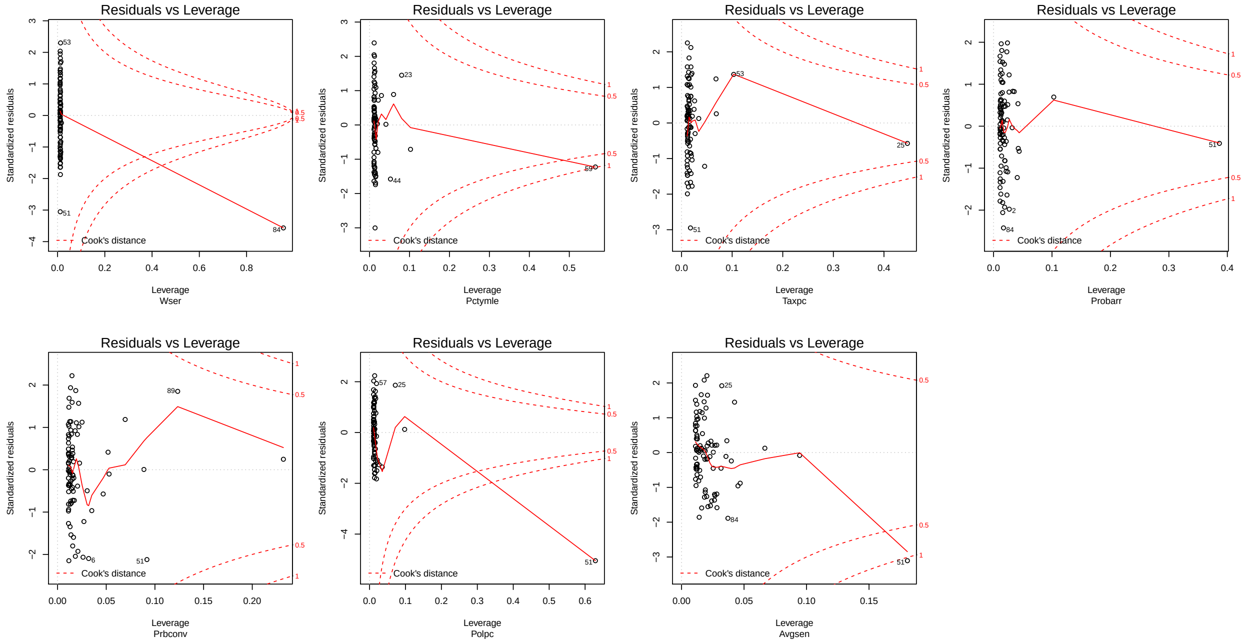


Figure 7: Cooks Distance for Selected Variables

We can see that for wser and pcymle, there are observations that we need to investigate further. For taxpc, there is an observation that is outside of the mean and pulling influence, but it is not outside of the Cook's distance boundary of 0.5.

For pctymle we see observation 59 is right at the Cook's distance boundary of 1.0. Further investigation shows that this record is for county 133. Assuming the county code is equivalent to FIPS code, this county is "Onslow" and contains the military base, "Marine Corps Base Camp Lejeune". This would explain as to why the percentage of young males is much higher in this county than other counties. Our policy recomendations do not extend to military bases, as these are typically governed by military police. Since we do not want this county and its unique population to inform policy, we will remove it from the study. This decision reduces our sample size to 89.

For wser we see an observation that is very far outside of Cook's distance boundary, with an average weekly service wage for the county being extremely high at $2177. One reason for this could be a single individual skewing the average for the country; therefore, we will impute this value by replacing it with the mean of the other counties.

```
##    county    crmrte   prbarr  prbconv  prbpris avgsen      polpc   density
## 51    115 0.0055332 1.090910 1.500000 0.500000  20.70 0.00905433 0.3858093
## 59    133 0.0551287 0.266960 0.271947 0.334951   8.99 0.00154457 1.6500655
## 84    185 0.0108703 0.195266 2.121210 0.442857   5.38 0.00122210 0.3887588
##       taxpc west central urban pctmin80     wcon     wtuc     wtrd
## 51 28.19310    1       0     0  1.28365 204.2206 503.2351 217.4908
## 59 27.46926    0       0     0 26.38140 264.0406 318.9644 183.2609
## 84 40.82454    0       1     0 64.34820 226.8245 331.5650 167.3726
##        wfir      wser   wmfg   wfed   wsta   wloc        mix    pctymle
## 51 342.4658  245.2061 448.42 442.20 340.39 386.12 0.10000000 0.07253495
## 59 265.1232  230.6581 258.25 326.10 329.43 301.64 0.12176319 0.24871162
## 84 264.4231 2177.0681 247.72 381.33 367.25 300.13 0.04968944 0.07008217
```

For polpc we see observation 51 far outside of Cook's distance, this row is also showing as an outlier for average sentence (avgsen) and is also the row with a probability of arrest greater than 1 discussed earlier. Therefore we will remove this record from our study.

Correlation

The final step in the exploratory data analysis is to look more closely at the correlation between variables, especially those involving the crime rate. Figure 8 shows the correlation of variables to crime rate.

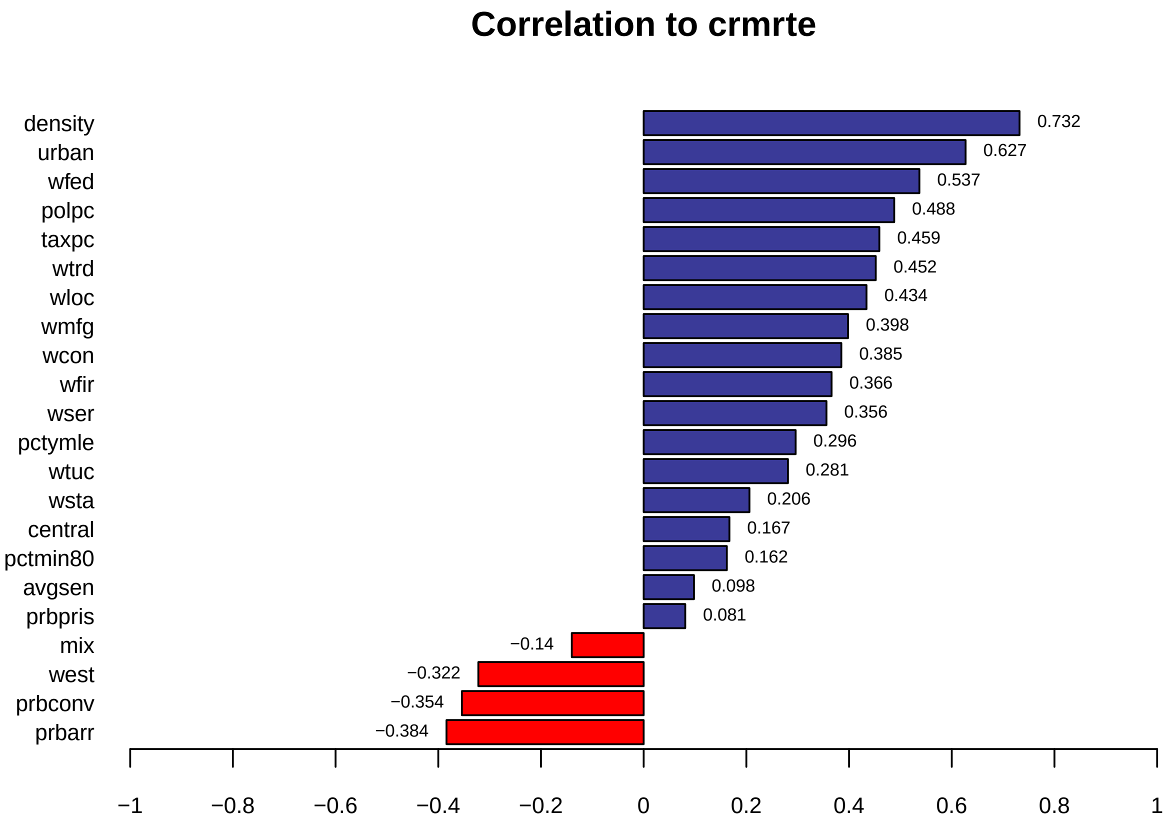


Figure 8: Correlation to Crime Rate

To understand the correlations between all variables in the dataset we construct the correlation matrix below in Figure 9, which illustrates correlation in a heatmap schema.

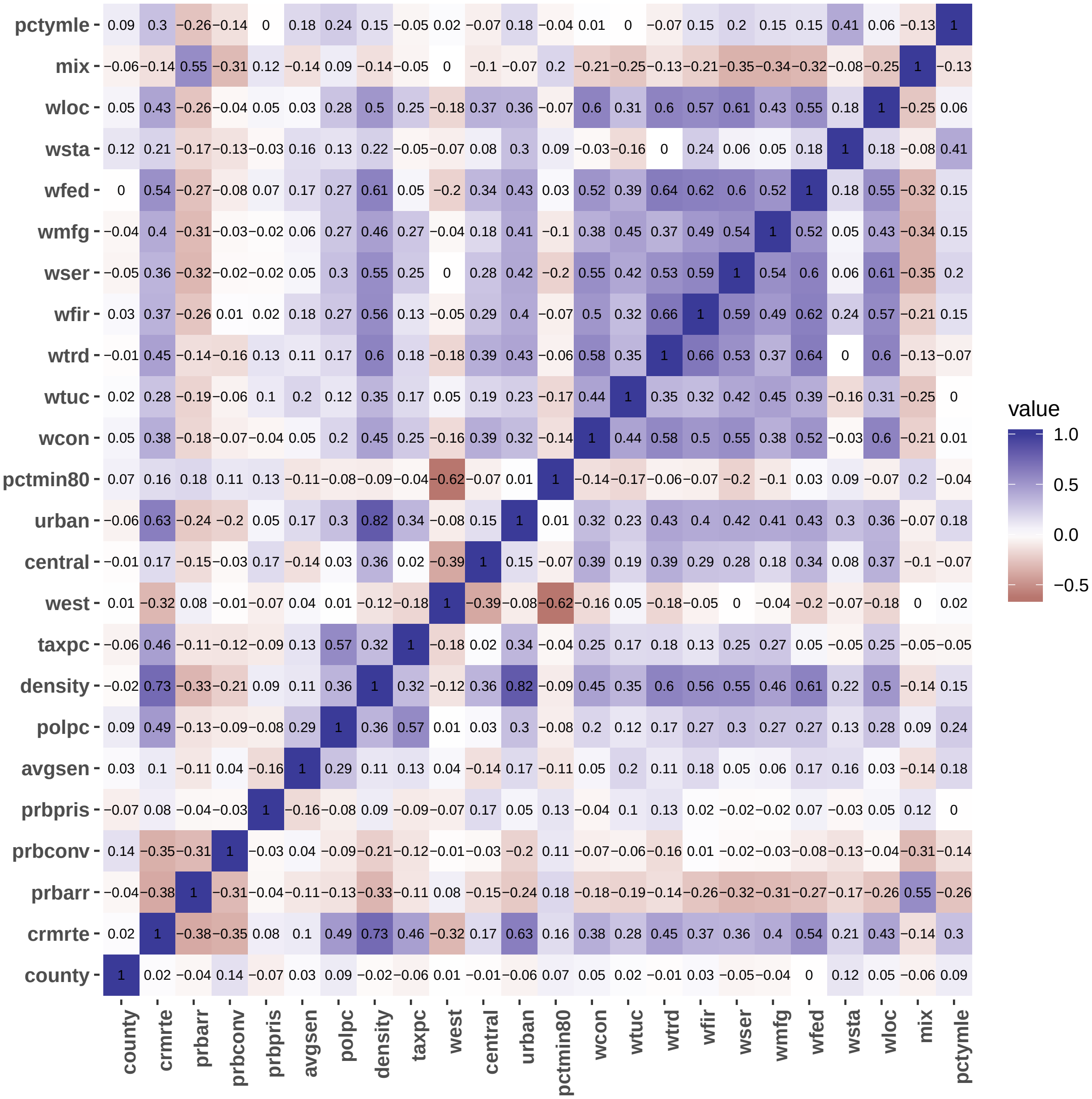


Figure 9: Correlation Chart

Other than year, we plot all variables with crmrte to observe their relationship visually. Year is not included as it a constant value and holds the same for all data points.

These charts will be useful when modelling, however we acknowledge that wser (weekly wage, service industry), mix, west, probability of arrest and probability of conviction are all negatively correlated with crime rate. Density has the largest positive correlation and is closely followed by urban. These variables help us to select our variables for Model 1.

Model 1

Measurements

This model is primarily focused on measuring the effect of density and the probability of conviction on crime. First we check the scatterplots of the two variables against the log of crmrte in Figure 10.

We can see from the scatterplots that using the log of crime rate is a good choice for the target variable, specifically for our initial analysis focusing on density which shows a positive linear relationship.

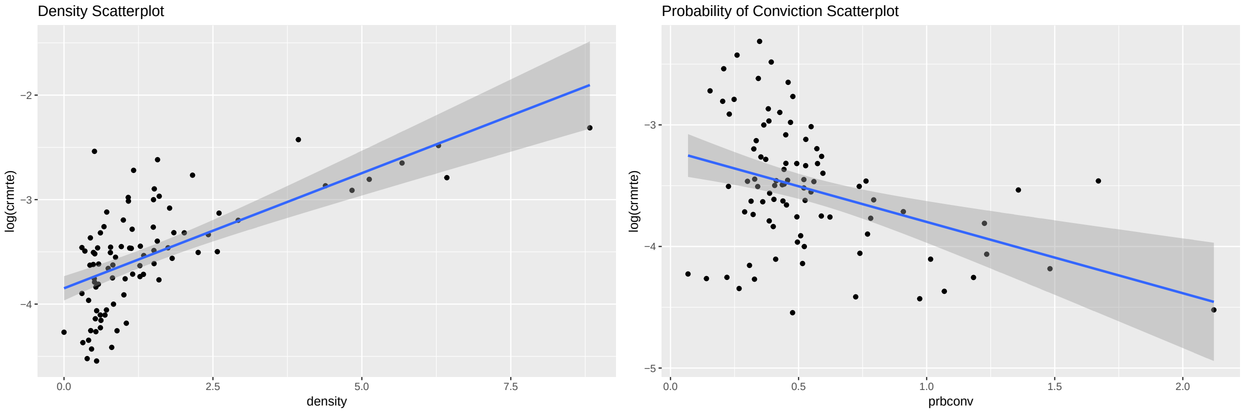


Figure 10: Model 1 Scatterplots

Covariate Selection

Based on the EDA, we will start building the first model using the variable for probability of conviction (prbconv) and including a covariate of density; these are the key variables of interest. Density has a strong positive correlation with crime rate (0.75). We also choose probability of conviction, as it has a relatively strong negative correlation with crmrte (-0.38) and we believe that generally the likelihood of being convicted is a stronger motivation against committing a crime then the probability of arrest, which is also negatively correlated with crime rate.

Transformations

No transformations are required outside predicting against log(crmrte).

Outliers

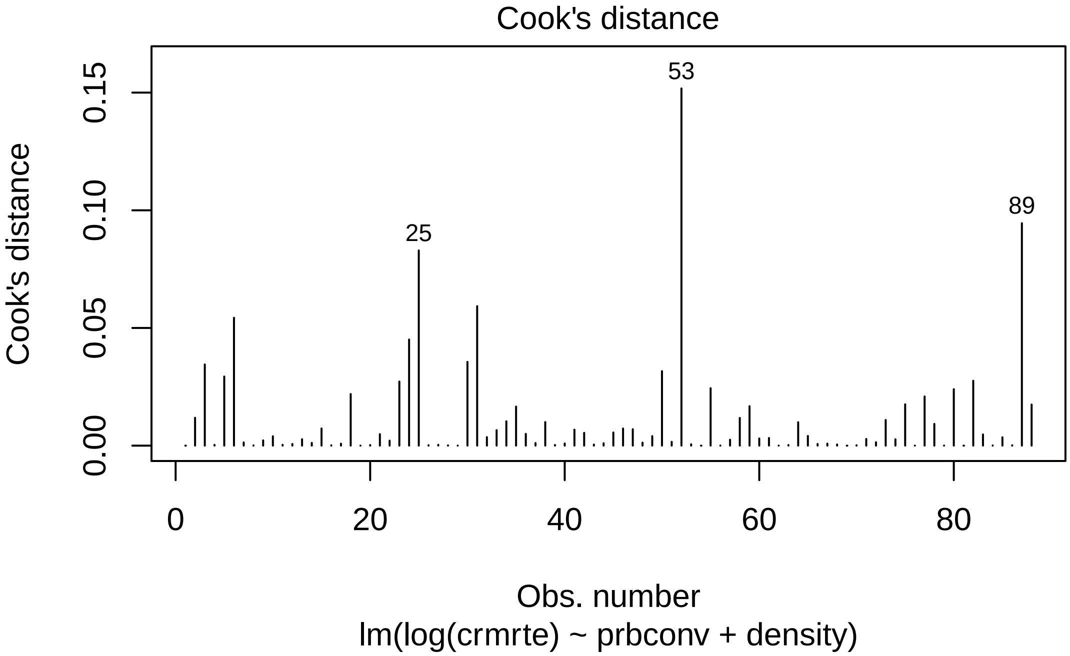


Figure 11: Model 1 Cooks Distance

The Cook's Distance Chart shows that no data point is outside the Cook's distance boundary of 0.5, which means we do not need to be concerned about outliers.

Assumptions

- **Assumption 1 (MLR1)**: The regression model is linear in the coefficients and the error term.
- **Assumption 2 (MLR2): Random Sampling**: We did not find evidence of clustering the sample.
- **Assumption 3 (MLR3): No perfect collinearity in independent variables** We can see from the earlier correlation chart that the correlation between density and prbconv is -0.23. In reality there is no need to explicitly check for perfect collinearity, because R will alert us if this rare condition happens.

However, We will use vif function to test for collinearity. We understand smallest possible value of VIF is one (absence of multicollinearity). As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity (James et al. 2014).

```
##  prbconv  density
## 1.048355 1.048355
```

The vif test showed low values which means no obvious collinearity.

- **Assumption 4 (MLR4): Zero Conditional Mean / Exogeneity**

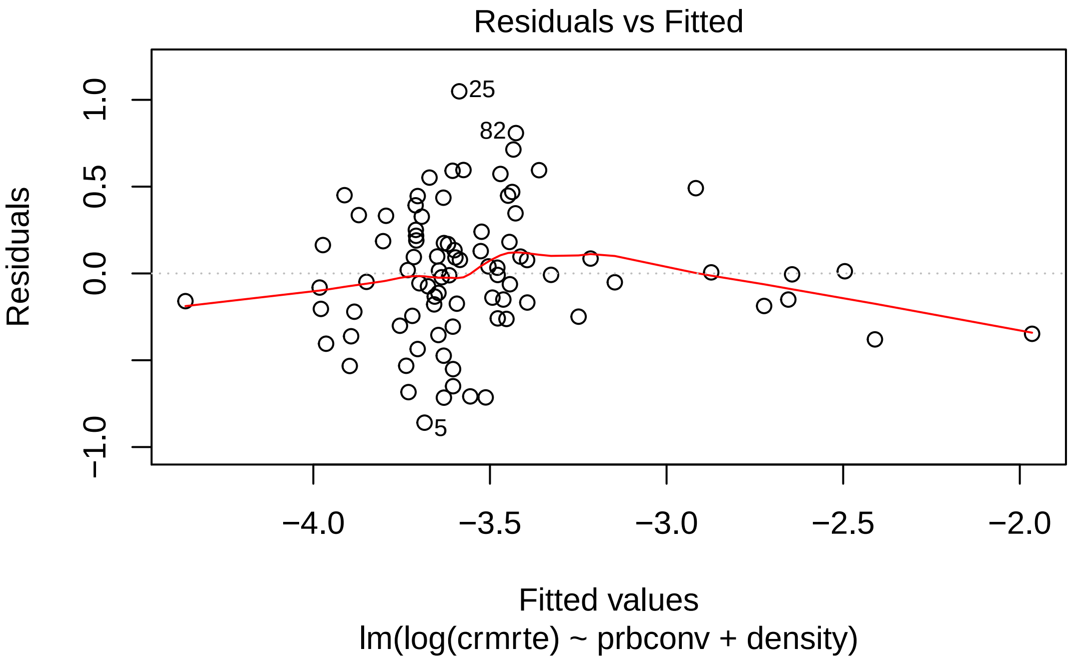


Figure 12: Model 1 Residuals vs Fitted

– In the Residual vs Fitted plot, the red line is around zero on the left but starts to go down on the right. Given our dataset has more data points on the left and less data on the right, this could explain the downslope on the right. Considering the lack of covariates, this could also be an indication of some information being held in the error term. As MLR assumptions 1-3 hold true, we can accept the weaker assumption of exogeneity.

- **Assumption 5 (MLR5): The error term has a constant variance (homoscedasticity)**

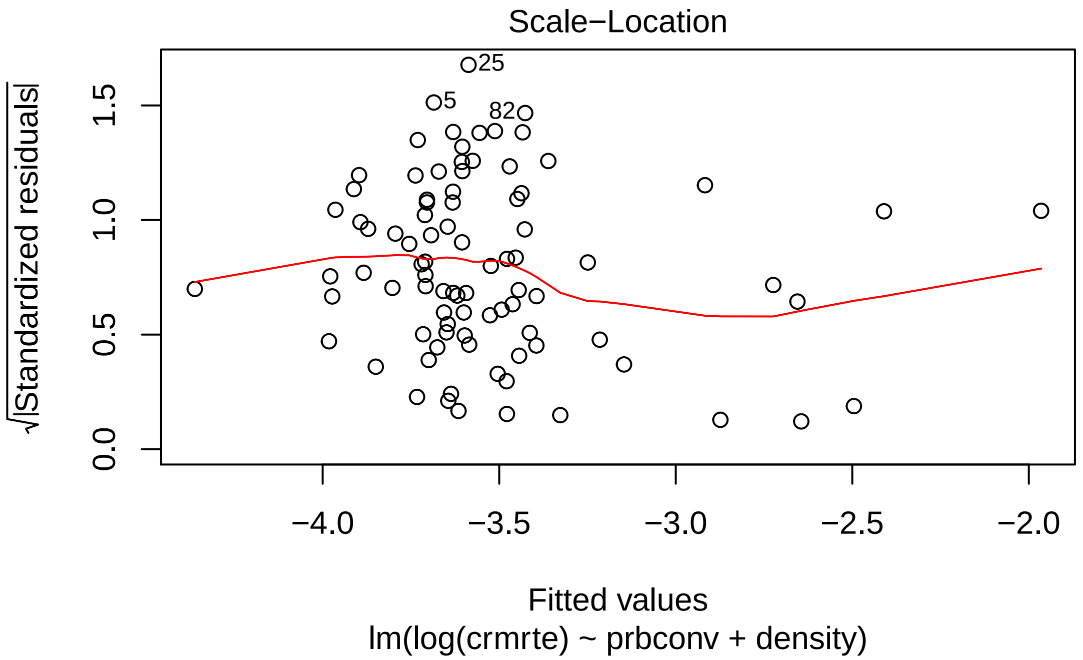


Figure 13: Model 1 Scale - Location

– The Residuals vs Fitted plot is wider at the left than the right but generally even. – The Scale-Location plot is not generally flat, but fewer data points were on the right, therefore we choose to conduct the Breusch-Pagan test.

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 7.524, df = 2, p-value = 0.02324
```

– The Breusch-Pagan test shows a p-value of 0.02324, which is statistically significant at 5% level, meaning we reject the null hypothesis (homoscedasticity).

– Homoscedasticity is a strong and relatively unrealistic assumption, we will use the more robust White standard errors in our analysis.

- **Assumption 6 (MLR6): Normality of error terms**

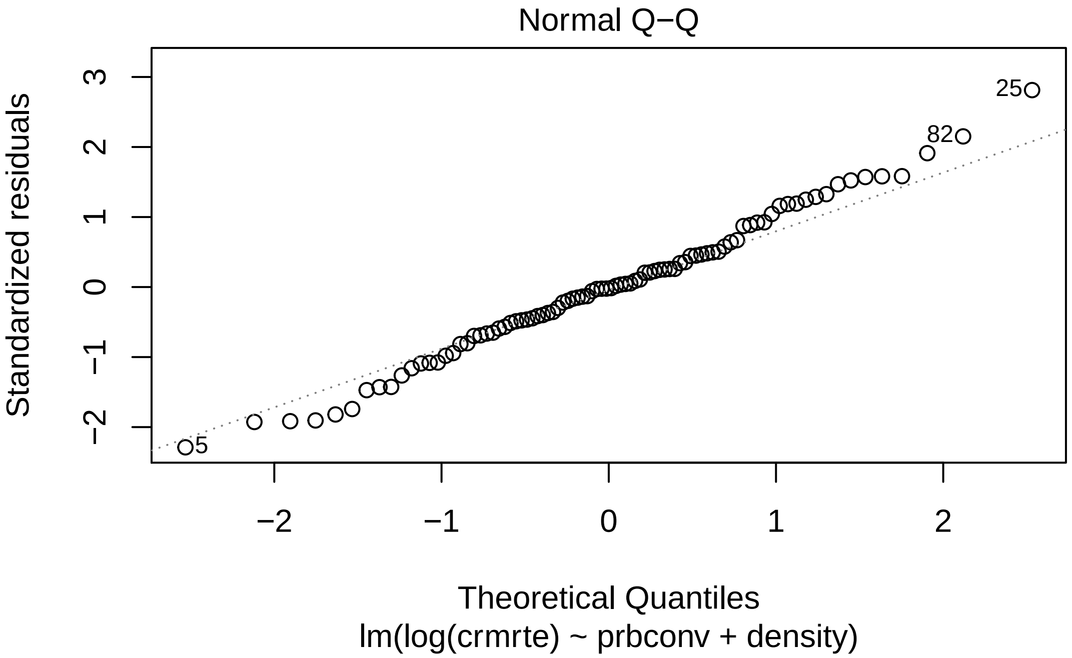


Figure 14: Model 1 Normal Q-Q

– The Q-Q plot shows unobserved errors are normally distributed. – We check with the Shapiro-Wilk normality test for confirmation

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.99214, p-value = 0.8821
```

– The p-value from the Shapiro-Wilk normality test is 0.8821, which is not statistically significant. In this case we are unable to reject the null hypothesis of the residuals having a normal distribution. – We can also check this conclusion with a histogram in Figuire 18.

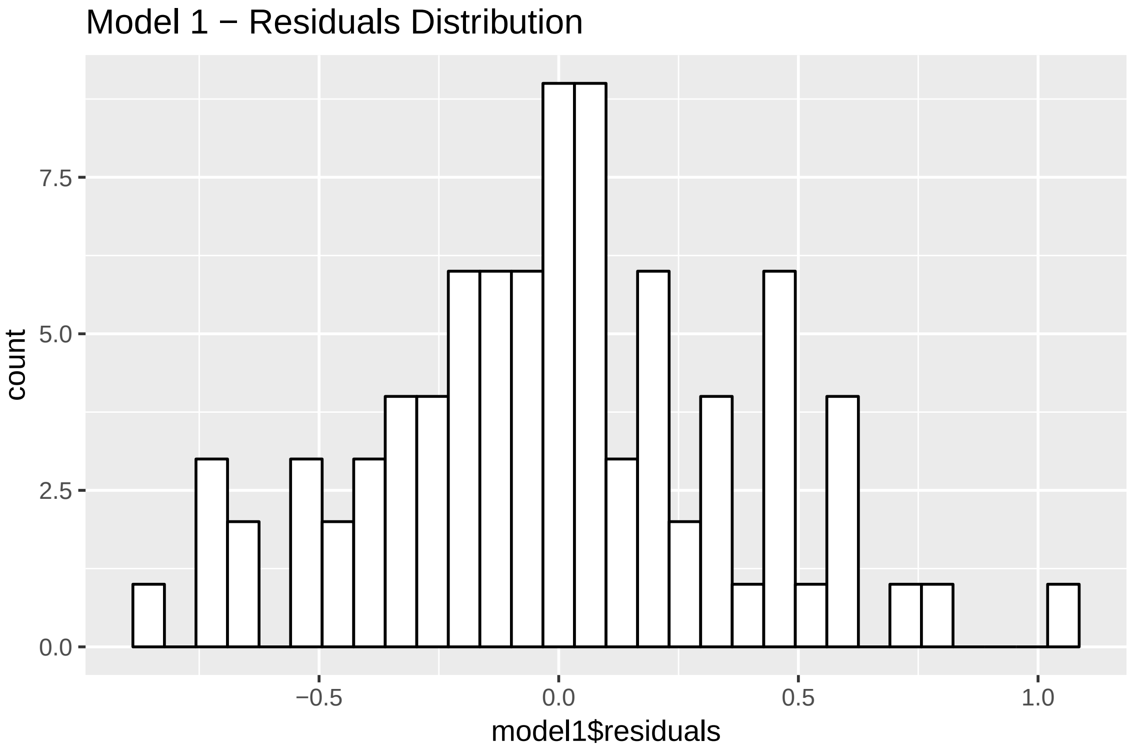


Figure 15: Model 1 Histogram of Residuals

– The Shapiro-Wilk test result, Q-Q plot and histogram all indicate there is no violation of normality of error terms assumption.

Model 1 Interpretation

- Generally MLR assumptions 1-6 were satisfied. There is an issue with MLR 4 where there may be some information held in the error term, so we will explore adding covariates in Models 2 and 3 to overcome this. We address the homoscedasticity by using more robust White standard error terms for MLR 5.

Covariates and r squared

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.608531   0.113624 -31.7587 < 2.2e-16 ***
## prbconv     -0.392193   0.128944  -3.0416  0.003129 **
## density      0.201608   0.027348   7.3720 1.021e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can observe both prbconv and density are extremely significant. More specifically, prbconv is statistically significant at 1% level and density is significant at 0.1% level.

```
## [1] "Adjusted r-squared:  0.472136366046685"
```

```
## [1] "AIC =  83.7722531634132"
```

We find a single point increase in the probability of convictions will have a 39% decrease in crime rate. An increase of 1,000 people per mile rise in density means an additional 20% increase in crime rate. Density has a greater impact than prbconv. The adjusted r-sqared is 0.47 which is equivalent to medium practical significance.

Predicting crmrte When log(crmrte) Is the Dependent Variable

The team realizes that when predicting crmrte using log(crmrte) as the dependent variable, we will need to avoid underestimating the expected value of crmrte. According to Woodridge book (Chapter 6), we will need to use the adjusted prediction of crmrte:

$$cr\hat{m}rte = exp(\hat{\sigma}^2/2)exp[log(cr\hat{m}rte)]$$

Where $\hat{\sigma^2}$ is simply the unbiased estimator of σ^2. Because σ^2, the standard error of the regression, is always reported, obtaining predicted values for crmrte is straight forward.

Adjusted r-squared: The model explains 47.2% of the variance of crime rate, which is medium, and suggests that there are more omitted variables being held in the residuals. We will investigate further with Models 2 and 3.

Omitted Variable Bias

Assuming the true data generating process for the model is described as follows:

$$log(crmrte) = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot prbconv + \mu$$

We will examine a possible bias by omitting the probablity of conviction (prbconv) from the model, by investigating when the estimated regression model does not include prbconv as a regressor.

First we check the correlation between density and probconv.

```
## [1] -0.2147671
```

The fact the correlation is -0.2147671 is cause for concern that omitting prbconv leads to a negatively biased estimate of $\hat{\beta_1}$. As a consequence we expect $\hat{\beta_1}$, the coefficient on density, to be too large in absolute value. Put differently, the OLS estimate of $\hat{\beta_1}$ suggests that a greater density will increase crime rate, but the effect of a greater density is overestimated, as it captures the effect of having higher probability of conviction.

To understand the magnitude of $\hat{\beta_1}$, we remove the variable prbconv to the regression and estimate the restricted model.

$$log(crmrte) = \beta_0 + \beta_1 \cdot density + \mu$$

We estimate both regression models and compare.

Table 1: Linear Models Predicting Crime Rate

	Dependent variable:	
	log(crmrte)	
	(1)	(2)
prbconv	−0.392	
density	0.202	0.220
Constant	−3.609	−3.849
AIC	83.772	91.946
Observations	88	88
R^2	0.484	0.421
Adjusted R^2	0.472	0.414
Residual Std. Error	0.379 (df = 85)	0.399 (df = 86)

We found the outcomes to be consistent with our expectations (density being overestimated with omitted variable bias, $0.220 > 0.202$).

Following the same process, we find prbconv to be underestimated with omitted variable bias.

Also, AIC indicates the first model (AIC = 83.772) to be a more efficient model.

Model 2

Measurements, Forward Variable Selection

The second model will include additional covariates to the explanatory variables of key interest selected in Model 1. The model further explores variables relating to geography (where counties are located) and demographics, specifically the percent of young males. We also examine other policing policy variables: probability of arrest and average sentence. Finally, we test a hypothesis on the coefficient of wages using an F-test.

Below is a summary of steps and tests we preformed:
1. Geographical variables (west/central) - F-test
2. Geographical variable (urban) - T-test
3. Legal variables (prbarr, prbconv, prbpris, avgsen) - F-test
4. Income variables (public sector: wfed, wsta, wloc) - F-test
5. Income variables (private sector: wtuc, wtrd, wfir, wser, wmfg) - F-test
6. Individual variables: taxpc, pctymle - T-tests for each variable

The following process describes a Forward Variable Selection Method:

We start with the dummy variables west and central.

For the first test, we would like to know whether a family could move to a region with lower crime rate.

Below is the model:

$$log(crmrte) = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot west + \beta_3 \cdot central + \mu$$

The restricted model would be

$$log(crmrte) = \beta_0 + \beta_1 \cdot density + \mu$$

We then conduct F test for the joint significance on "west" and "central"

$$H_0 = \beta_{west} = 0, \beta_{central} = 0; \quad H_a = \beta_{west} \neq 0, \beta_{central} \neq 0$$

```
## Linear hypothesis test
##
## Hypothesis:
## west = 0
## central = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ density + west + central
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     86
## 2     84  2 12.037 2.532e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.669007   0.074682 -49.1281 < 2.2e-16 ***
## density      0.229521   0.032342   7.0967 3.763e-10 ***
## west        -0.454075   0.092648  -4.9011 4.585e-06 ***
## central     -0.218864   0.097362  -2.2479    0.0272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test shows statistically significant with a very low Pr(>F), and we are able to rejct the null hypothesis.

Now we want to know whether moving out of urban will allow a family to relocate to a lower crime rate area. We are going to test statistical significance on single variable urban, using white standard error.

$$log(crmrte) = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot west + \beta_3 \cdot central + \beta_4 \cdot urban + \mu$$

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) -3.691833   0.084086 -43.9054 < 2.2e-16 ***
## density      0.266776   0.065361   4.0816 0.0001024 ***
## west        -0.460665   0.097098  -4.7443 8.603e-06 ***
## central     -0.242079   0.101303  -2.3896 0.0191304 *
## urban       -0.225189   0.316844  -0.7107 0.4792477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value for the single variable urban is 0.479, and we are not able to reject the null hypothesis of the coefficient of urban being zero at 5% level.

As we also suspect the wage could have influence crime rate, so we decide to test the joint significance of wages. We split the wages into 2 groups - private sector and public sector:

- Group 1: wcon, wtuc, wtrd, wfir, wser, wmfg
- Group 2: wfed, wsta, wloc

Since these variables are based on wages (currency), we decided to use logarithm transformation, according to recommended good practice described by Woodridge. The outputs of the analysis are as follows:

For Group 1, the DF 6 F-value is 1.0536, and the p-value is 0.3977 This means we are unable to reject null hypothesis at 5% level ($H_0 = \beta_{log(wcon)} = 0, \beta_{log(wtuc)} = 0, \beta_{log(wtrd)} = 0, \beta_{log(wser)} = 0, \beta_{log(wmfg)} = 0; \quad H_a = \beta_{log(wcon)} \neq 0, \beta_{log(wtuc)} \neq 0, \beta_{log(wtrd)} \neq 0, \beta_{log(wser)} \neq 0, \beta_{log(wmfg)} \neq 0$).

Group 1

```
## Linear hypothesis test
##
## Hypothesis:
## log(wcon) = 0
## log(wtuc) = 0
## log(wtrd) = 0
## log(wfir) = 0
## log(wser) = 0
## log(wmfg) = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ density + west + central + log(wcon) + log(wtuc) +
##     log(wtrd) + log(wfir) + log(wser) + log(wmfg)
##
## Note: Coefficient covariance matrix supplied.
```

```
##
##   Res.Df Df      F Pr(>F)
## 1     84
## 2     78  6 1.0536 0.3977
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -7.494472   2.472240 -3.0315  0.003302 **
## density      0.194150   0.038753  5.0099 3.315e-06 ***
## west        -0.450717   0.100073 -4.5039 2.307e-05 ***
## central     -0.257622   0.110347 -2.3347  0.022135 *
## log(wcon)    0.251084   0.395337  0.6351  0.527213
## log(wtuc)    0.151509   0.303186  0.4997  0.618677
## log(wtrd)    0.210810   0.570794  0.3693  0.712884
## log(wfir)   -0.212216   0.503342 -0.4216  0.674467
## log(wser)   -0.102648   0.378642 -0.2711  0.787034
## log(wmfg)    0.385089   0.236974  1.6250  0.108192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For Group 2, the DF 3 F-value is 3.2091, and the p-value is 0.02733 This means Group 2 is also jointly significant at 5% level ($H_0 = \beta_{log(wfed)} = 0, \beta_{log(wsta)} = 0, \beta_{log(wloc)} = 0; \quad H_a = \beta_{log(wfed)} \neq 0, \beta_{log(wsta)} \neq 0, \beta_{log(wloc)} \neq 0$)

Group 2

```
## Linear hypothesis test
##
## Hypothesis:
## log(wfed) = 0
## log(wsta) = 0
## log(wloc) = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ density + west + central + log(wfed) + log(wsta) +
##     log(wloc)
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1     84
## 2     81  3 3.2091 0.02733 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -12.2139052   3.9010447 -3.1309  0.002424 **
## density       0.1718887   0.0345686  4.9724 3.636e-06 ***
## west         -0.4202458   0.0922092 -4.5575 1.811e-05 ***
## central      -0.2661210   0.1015389 -2.6209  0.010471 *
## log(wfed)     1.0470145   0.5378570  1.9466  0.055044 .
## log(wsta)    -0.0026191   0.3472370 -0.0075  0.994000
## log(wloc)     0.3975025   0.7273738  0.5465  0.586231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With F value 3.2091 and P value 0.02733, we are able to reject the null hypothesis which is:

$$\beta_{log(wfed)} = 0, \beta_{log(wsta)} = 0, \beta_{log(wloc)} = 0$$

Therfore Model 2 can now be defined as:

$$log(crmrte) = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot west + \beta_3 \cdot central + \beta_4 \cdot log(wfed) + \beta_5 \cdot log(wsta) + \beta_6 \cdot log(wloc) + \mu$$

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) -12.2139052   3.9010447 -3.1309  0.002424 **
## density       0.1718887   0.0345686  4.9724 3.636e-06 ***
## west         -0.4202458   0.0922092 -4.5575 1.811e-05 ***
## central      -0.2661210   0.1015389 -2.6209  0.010471 *
## log(wfed)     1.0470145   0.5378570  1.9466  0.055044 .
## log(wsta)    -0.0026191   0.3472370 -0.0075  0.994000
## log(wloc)     0.3975025   0.7273738  0.5465  0.586231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The legal variables we have not examined are:

- prbarr
- prbconv
- prbpris
- avgsen

Given these variables are all related to punishment, we decided to conduct a joint significance test:

$$\begin{gathered} log(crmrte) = \beta_0 + \beta_1 \cdot density + \beta_2 \cdot west + \beta_3 \cdot central + \\ \beta_4 \cdot log(wfed) + \beta_5 \cdot log(wsta) + \beta_6 \cdot log(wloc) + \\ \beta_7 \cdot prbarr + \beta_8 \cdot prbconv + \beta_9 \cdot prbpris + \beta_{10} \cdot avgsen + \mu \end{gathered}$$

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## prbpris = 0
## avgsen = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ density + west + central + log(wfed) + log(wsta) +
##     log(wloc) + prbarr + prbconv + prbpris + avgsen
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     81
## 2     77  4 8.7524 7.057e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -9.7388301  3.2758528 -2.9729 0.0039361 **
## density      0.1093768  0.0266015  4.1117 9.741e-05 ***
## west        -0.4271657  0.0777330 -5.4953 4.864e-07 ***
## central     -0.2649003  0.0756176 -3.5032 0.0007691 ***
## log(wfed)    1.0490289  0.4641002  2.2603 0.0266248 *
## log(wsta)   -0.2839724  0.2907171 -0.9768 0.3317272
## log(wloc)    0.4035547  0.5984711  0.6743 0.5021340
## prbarr      -1.5468939  0.3860065 -4.0074 0.0001406 ***
## prbconv     -0.6347074  0.1315379 -4.8253 6.924e-06 ***
## prbpris      0.0774368  0.4463597  0.1735 0.8627253
## avgsen      -0.0022634  0.0149989 -0.1509 0.8804457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With DF 4, the F value is 8.7524, which is very significant.

While examining the coefficients, we noticed the coeffienct of prbpris is 0.0774368, a positive number. This is counter intuitive as we expected the higher 'probability' of prison sentence would lead to a reduced crime rate instead of a increased crime rate. We decided to conduct another hypothesis test to see whether without prbpris, the group of variables (prbarr, prbconv and avgsen) are still significant:

```
## Linear hypothesis test
##
## Hypothesis:
## prbarr = 0
## prbconv = 0
## avgsen = 0
##
## Model 1: restricted model
## Model 2: log(crmrte) ~ density + west + central + log(wfed) + log(wsta) +
##     log(wloc) + prbarr + prbconv + avgsen
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1     81
## 2     78  3 11.739 1.992e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -9.691880   3.239368 -2.9919   0.00371 **
## density      0.109566   0.026484  4.1370 8.808e-05 ***
## west        -0.427368   0.075343 -5.6723 2.291e-07 ***
## central      -0.263335   0.077147 -3.4134   0.00102 **
## log(wfed)     1.051333   0.462552  2.2729   0.02578 *
## log(wsta)    -0.285051   0.286004 -0.9967   0.32201
## log(wloc)     0.400164   0.599183  0.6678   0.50620
## prbarr       -1.548473   0.376804 -4.1095 9.717e-05 ***
## prbconv      -0.635108   0.129473 -4.9053 4.989e-06 ***
## avgsen       -0.002635   0.015199 -0.1734   0.86281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F test shows it is still significant even at 1% significant level with p-value 1.992e-06.

We then conducted similar t tests for other variables with null hypothesis of $\beta = 0$ and alternative hypothesis $\beta \neq 0$:

- taxpc (whether increasing/reducing tax will reduce crime rate). T value is 0.0546, and p value is 0.956572

- pctymle (whether having higher density of young male could increase crime rate). T value is 0.5014, and p value is 0.6175289

We are able to reject null hypothesis of $\beta_{pctymle} = 0$ at 5% level and also reject null hypothesis of $\beta_{taxpc} = 0$ at 5% significant level.

Now we are going to examine the AIC to identify the efficience of our model:

```
## [1] "AIC =  32.9273887931797"
```

The following charts in Figure 16 show the additional selected coefficients to Model 1.

Model 2: log(crmrte) ~ density + west + central + log(wfed) + log(wsta) + log(wloc) + prbarr + prbconv + avgsen

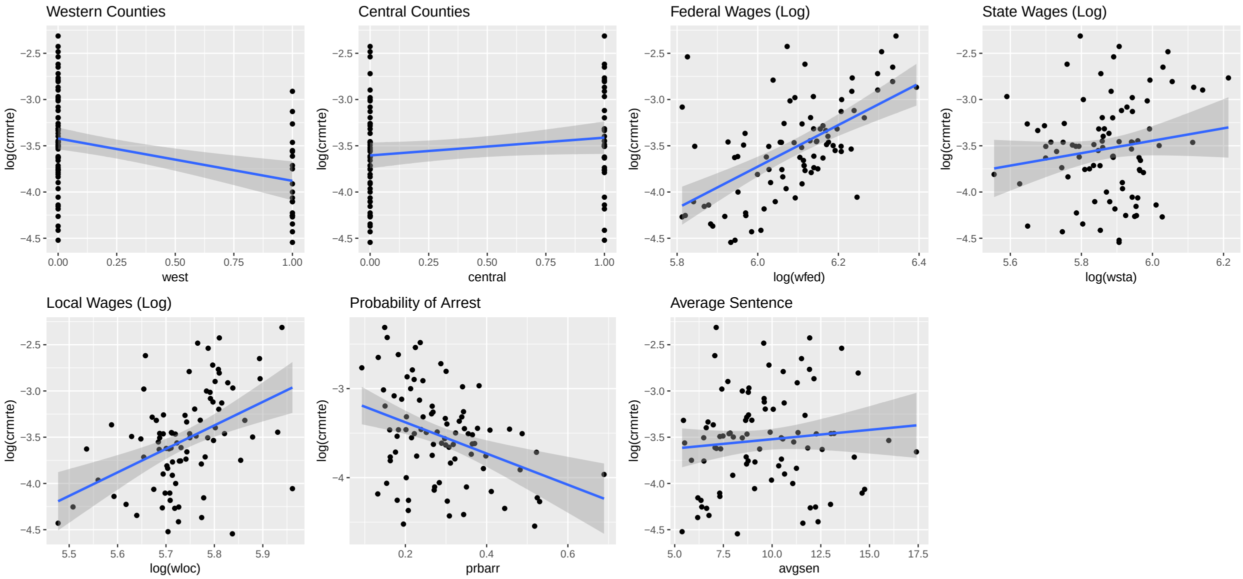


Figure 16: Model 2 Additional Scatterplots

Outliers

Given our chosen model, we check for outliers.

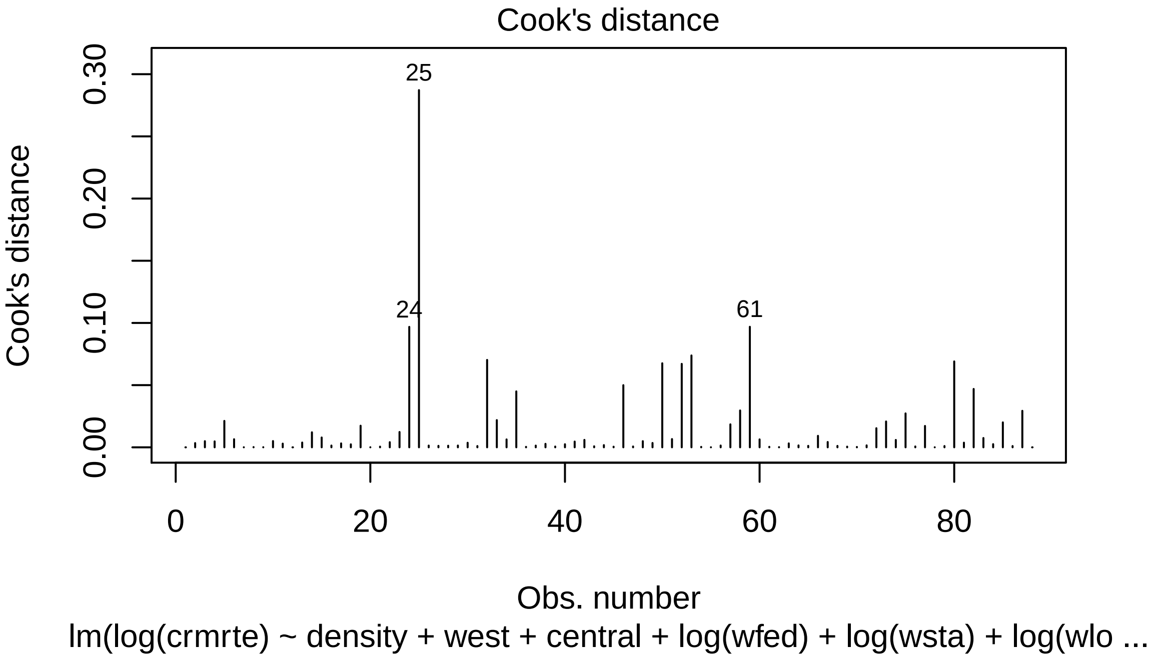


Figure 17: Model 2 Cooks Distance

- The Cook's distance chart shows data points with a Cook's distance greater than 1.0 or approaching 0.5, which means we do need to be concerned about outliers.

We move on to checking the classical linear model assumptions.

Assumptions

- **Assumption 1 (MLR1)**: The regression model is linear in the coefficients and the error term, complying with assumption 1.
- **Assumption 2 (MLR2): Random Sampling**: All variables taken from the census should accurately reflect the population of the counties and be IID, as the census is sent to every individual. Counts taken from the FBI reports should include every arrest, so no sampling took place. It is unknown how the wage variables were created from the North Carolina Employment Security Commission, but for this analysis, we assume the wages samples to create the average wage were IID. Model 2 does not violate assumption 2.
- **Assumption 3 (MLR3): No perfect collinearity in independent variables**

We will use vif function to test for collinearity:

```
##   density      west   central log(wfed) log(wsta) log(wloc)    prbarr
##  1.940290  1.195923  1.448904  1.769789  1.106864  1.562657  1.394183
##   prbconv    avgsen
##  1.299152  1.114031
```

The vif test showed low values which means there is no cause for concern.

- **Assumption 4 (MLR4): Zero Conditional Mean / Exogeneity**

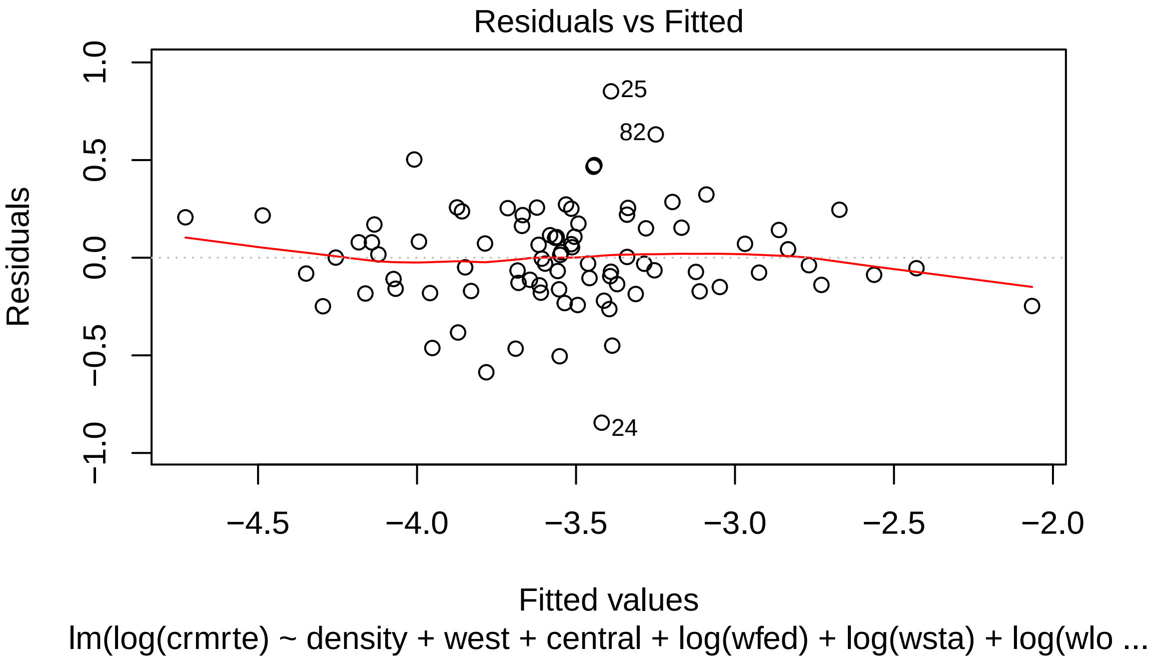


Figure 18: Model 2 Residuals Vs Fitted

– In the Residual vs Fitted plot, the red line of best fit generally holds around zero starting at the minimum fitted values. There is a slight negative deviation of the mean away from zero for the more positively fitted values, but given that our dataset has fewer data points for the rightmost values, the negative deviation of the mean from zero does not singularly provide enough evidence to violate assumption 4.

- **Assumption 5 (MLR5): The error term has a constant variance (homoscedasticity)**

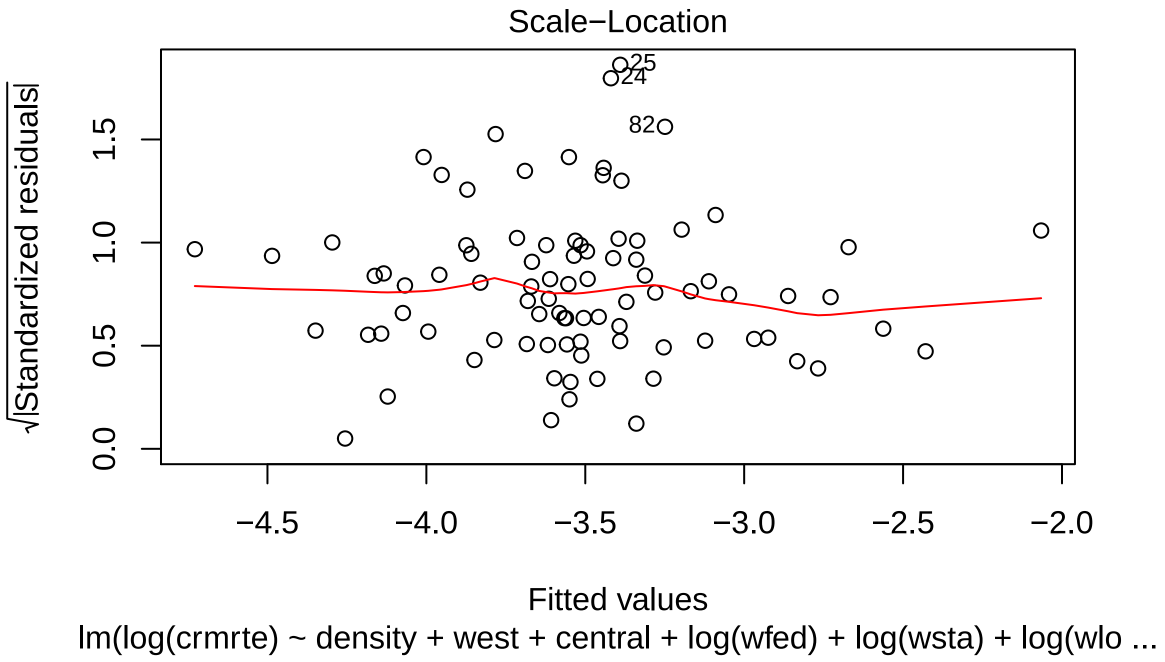


Figure 19: Model 2 Scale - Location

– The Residuals vs Fitted plot does not fan out with the exception of a few of the rightmost fitted values – To explore homoskedasticity, we can view the moving mean of the Scale-Location plot. Although the line is relatively flat towards the ends, the team recognizes there are a few inflection points in the middle of the graph that could reflect heteroskedasticity. To further investigate Model 2's adherence to assumption 5, we conduct the Breusch-Pagan test to measure if the variance of residuals is independent of the values of the covariates.

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 25.96, df = 9, p-value = 0.002074
```

– The Breusch-Pagan test shows a statistically significant p-value of 0.002074, which means we are able to reject the null hypothesis (homoscedasticity) at 5% level. Although this provides some evidence of Model 2 violating assumption 5, we will use the more robust White standard errors in our analysis.

- **Assumption 6 (MLR6): Normality of error terms**

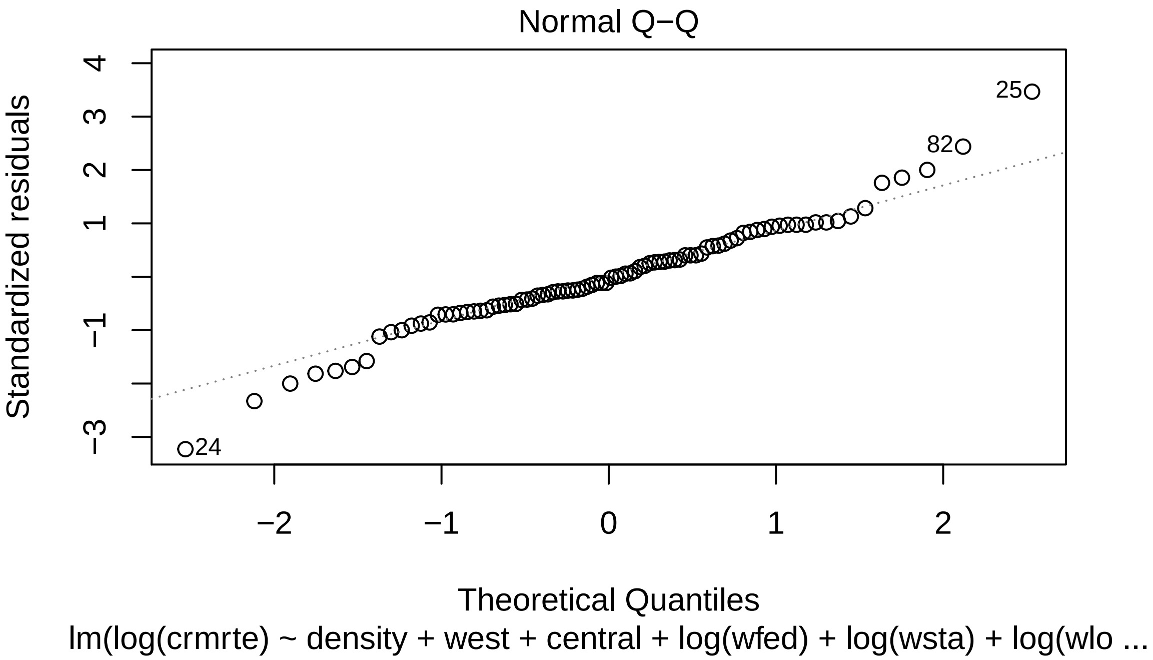


Figure 20: Model 2 Normal Q-Q

– To check if Model 2 adheres to assumption 6, the research team first observes the Q-Q plot to explore if the residuals are relatively normally distributed. We see most of the observed instances remain close to the regression line. As some of the leftmost instances stray from the line, the team will run the Shapiro-Wilk normality test:

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.9719, p-value = 0.05262
```

The p-value from the Shapiro-Wilk normality test is 0.05262, which is not quite statistically significant. In this case we are unable to reject the null hypothesis (residuals having a normal distribution) at 5% significant level. However, we continue to explore the null hypothesis of the residuals through a histogram of the residuals. In this case, CLT can still be applied and we can assume residual is part of a normal distribution.

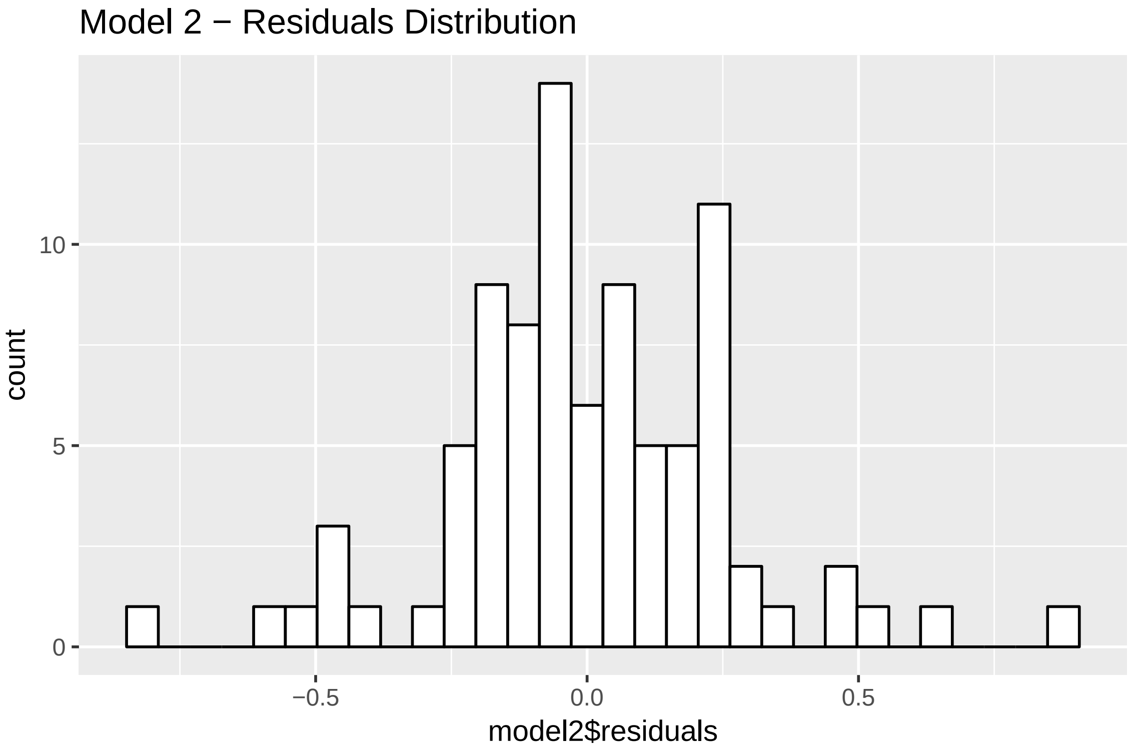


Figure 21: Model 2 Histogram of Residuals

Model 2 Interpretation

- Generally MLR assumptions 1-6 were satisfied. We now check the statistical impact of the variables selected and the practical significance of using these in the model.

Covariates and r squared

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -9.691880   3.239368 -2.9919   0.00371 **
## density      0.109566   0.026484  4.1370 8.808e-05 ***
## west        -0.427368   0.075343 -5.6723 2.291e-07 ***
## central     -0.263335   0.077147 -3.4134   0.00102 **
## log(wfed)    1.051333   0.462552  2.2729   0.02578 *
## log(wsta)   -0.285051   0.286004 -0.9967   0.32201
## log(wloc)    0.400164   0.599183  0.6678   0.50620
## prbarr      -1.548473   0.376804 -4.1095 9.717e-05 ***
## prbconv     -0.635108   0.129473 -4.9053 4.989e-06 ***
## avgsen      -0.002635   0.015199 -0.1734   0.86281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "adjusted r-squared:  0.724687462491743"
```

```
## [1] "AIC =  32.9273887931797"
```

The first thing to note is that the average sentence coefficient does not have statistical significance on the model, but our intuition around location and policing policy seem to hold true.

For each coefficient we discuss the impact:

- **density** For each increase of 1,000 people per mile, there is an additional 10.9% in crime rate.

- **west** A county in the western region could have 42.7% decrease in crime rate.

- **central** A county in the central region could have 26.3% decrease in crime rate.

- **wfed** From the surface, a 1% increase in the weekly federal wage has a positive impact of 1.05% on crime rate.

- **wsta** From the surface, a 1% increase in the weekly federal wage has a negative impact of 0.28% on crime rate.

- **wloc** From the surface, a 1% increase in the weekly federal wage has a positive impact of 0.4% on crime rate.

- **prbarr** One point increase in the probability of arrests can cause 154.8% decrease on crime rate.

- **prbconv** One point increase in the probability of conviction can cause 63.5% decrease on crime rate.

- **avgsen** Every extra day in sentencing leads to 0.2% crime rate drop.

Adjusted r-squared: The model explains 72.5% of the variance of crime rate, which is practically significant (> 0.5).

We noticed log(wfed) and log(loc) have positive relationships with crime rate and that log(wsta) has a negative relationship. This suggests there might be an omitted variable that causes high crime rate and drives weekly federal/local rage higher (or state wage lower) at the same time. This omitted variable seems to have a positive correlation with crime rate (crmrte), and we belive it could lead to an overestimate or underestimate for known variables, depending on their correlation with this omitted variable.

Also, we consider there are other omitted variables in this analysis. For example, education level could potentially be negatively correlated with crime rate. Omitted education level could lead to overestimates (positive bias) for variables like density, west and central and underestimates for variables like prbarr, prbconv and avgsen.

Model 3

Measurements

The third model will include virtually all variables available, including variables related to wage, offence mix (face to face or other), urban classification, tax revenue and police per capital, which were purposefully excluded in earlier models. The goal of Model 3 is to help show the robustness of Model 2 and to highlight some risks in including too many variables.

Covariate Selection

To compare the model, we used same variables in Model 2 and the remaining variables, except for year and county (year is same for all data points and county is a unique identifier for all observations).

Transformations

Per our discussion in earlier sections, we decided to use the logarithm for wages and level variables for all others.

Outliers

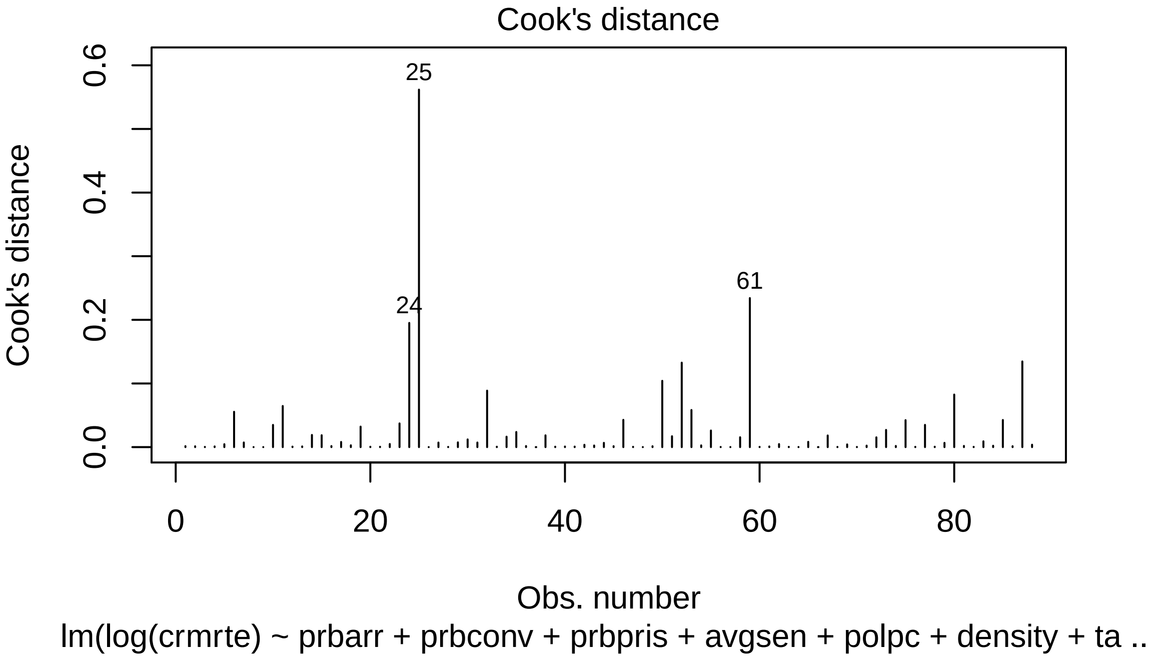


Figure 22: Model 3 Cook's Distance

The Cook's distance chart shows that there at least one data point with a Cook's distance greater than 0.5. Although an argument could be made to remove the outlier that is greater than 0.5, we decide to keep it to avoid overfitting.

Assumptions

- **Assumption 1 (MLR1)**: The regression model is linear in the coefficients and the error term.
- **Assumption 2 (MLR2): Random Sampling**: We did not find evidence of clustering the sample.
- **Assumption 3 (MLR3): No perfect collinearity in independent variables**

We will use vif function to test for collinearity:

```
##    prbarr   prbconv   prbpris    avgsen     polpc   density     taxpc
##  2.089181  1.677503  1.223844  1.516372  2.412983  5.523362  2.491211
##      west   central     urban  pctmin80 log(wcon) log(wtuc) log(wtrd)
##  3.330110  2.074790  4.034348  2.989499  2.151332  1.745053  3.106660
## log(wfir) log(wser) log(wmfg) log(wfed) log(wsta) log(wloc)       mix
##  2.569539  2.479917  2.217892  3.141964  1.737542  2.342461  2.247888
##   pctymle
##  1.622583
```

The vif test showed density has 5.52 which is greater than 5 and is a cause for concern. If we were to continue optimizing this model, we would need to remove the density variable which might lead to a simpler model without compromising the model accuracy.

- **Assumption 4 (MLR4): Zero Conditional Mean / Exogeneity**

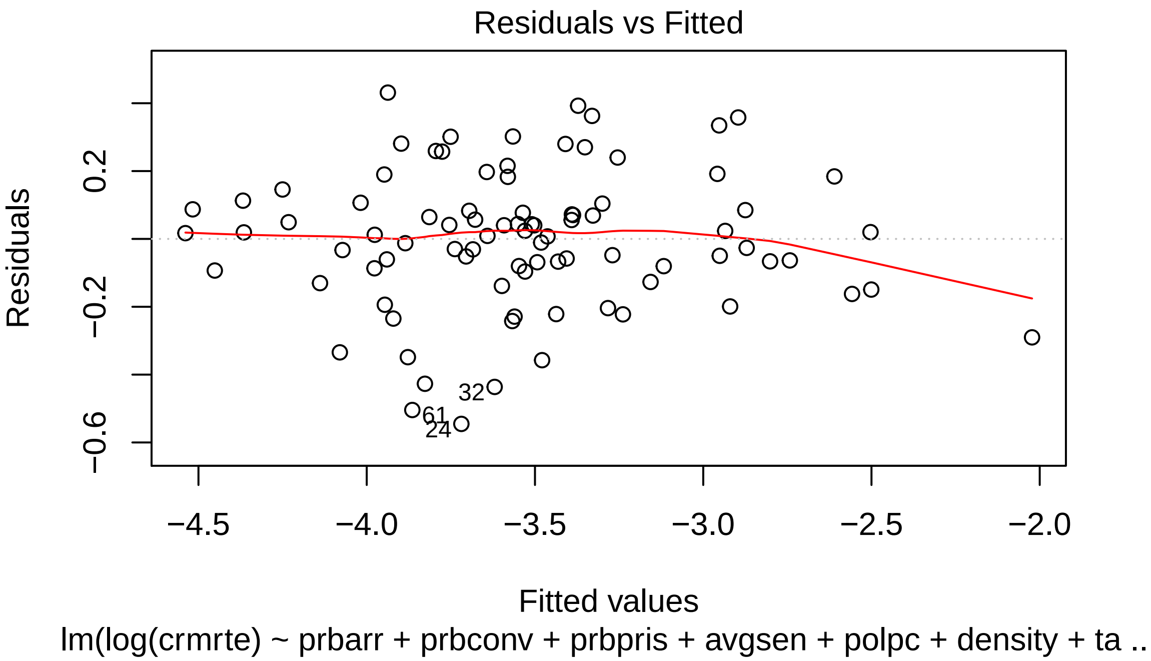


Figure 23: Model 3 Residuals Vs Fitted

– In the Residual vs Fitted plot, the red line of best fit does not hold around zero and skews sharply down and to the right. This model does violate assumption 4 and cannot be rectified by adding more variables, considering all variables are used in the model. We would need to further examine transforming some of the variables or to consider excluding some.

- **Assumption 5 (MLR5): The error term has a constant variance (homoscedasticity)**

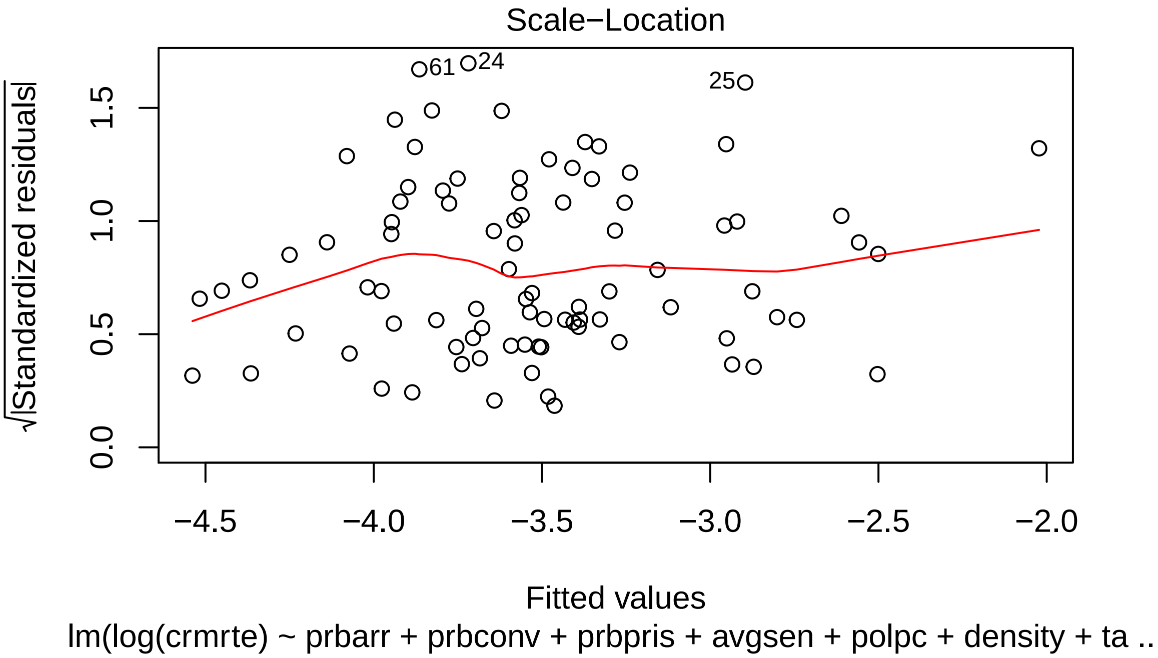


Figure 24: Model 3 Scale - Location

– To explore homoskedasticity, we can view the moving mean of the Scale-Location plot. The red line is not flat and is inconsistent through the data. The chart suggests that this model is in violation of assumption 5. We conduct the Breusch-Pagan test to measure if the variance of residuals is independent of the values of the covariates.

```
## [1] "bptest test:"
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 31.131, df = 22, p-value = 0.09347
```

A p-value of 0.09347 is not statistically significant at 5% level, and therefore we are not able to reject the null hypothesis of homoscedasticity at 5% significance level. We will still use White standard error for coeffienct.

- **Assumption 6 (MLR6): Normality of error terms**

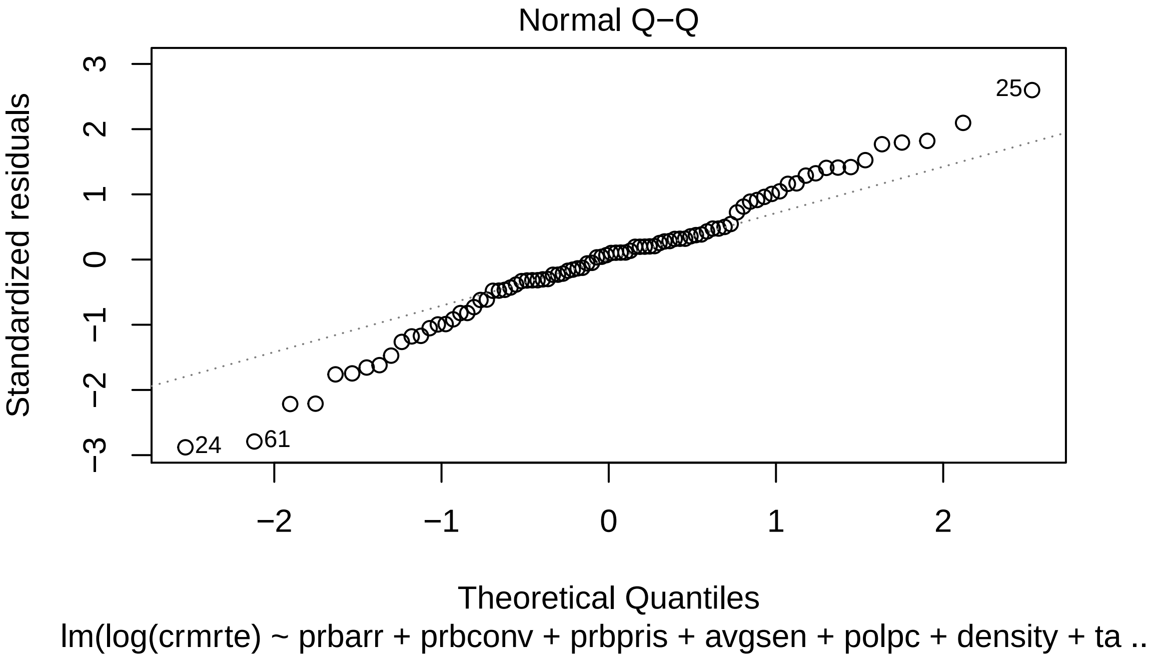


Figure 25: Model 3 Normal Q-Q

– To check if Model 3 adheres to assumption 6, the research team first observes the Q-Q plot to explore if the residuals are relatively normally distributed. We see most of the observed instances remain close to the regression line, however there is considerable deflection away from the line at the lower and upper ends, much more than in Models 1 and 2. The team runs the Shapiro-Wilk normality test:

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.98233, p-value = 0.2747
```

The p-value from the Shapiro-Wilk normality test is 0.2747, which is not statistically significant. In this case we are unable to reject the null hypothesis of residuals having a normal distribution. The research team explores the null hypothesis of the residuals having a normal distribution through a histogram of the residuals.

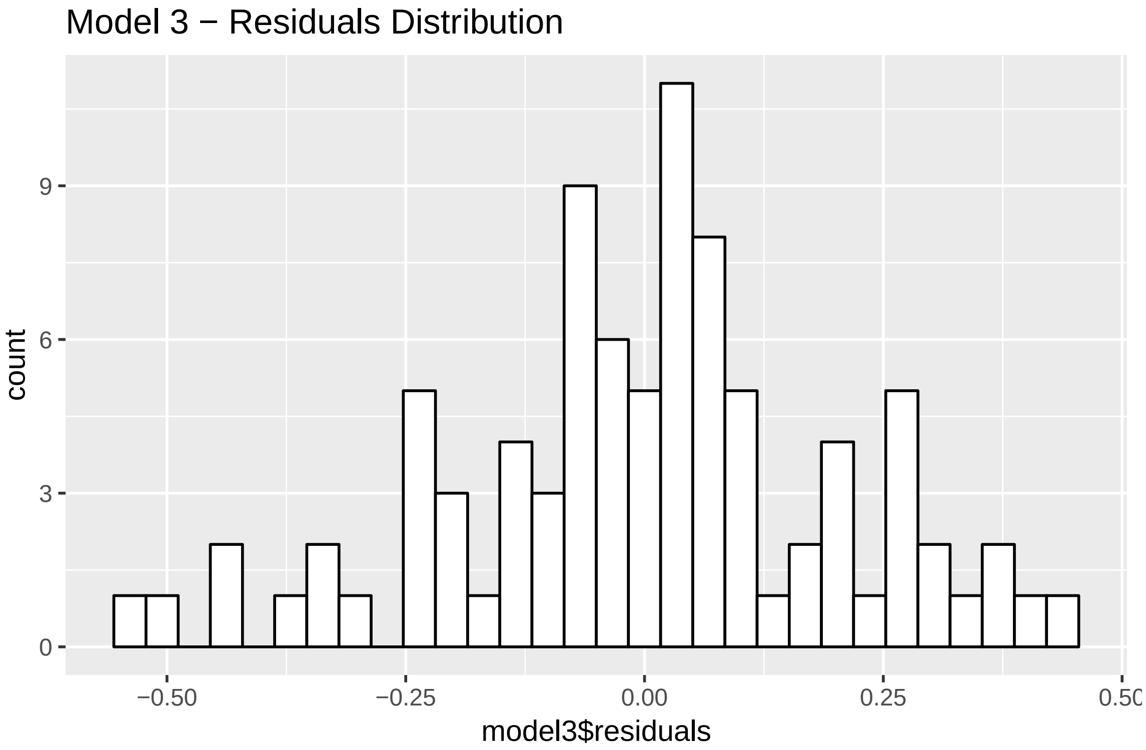


Figure 26: Model 2 Histogram of Residuals

The histogram does indicate a somewhat normal distribution and lines up with the Q-Q plot analysis. Given this and the Shapiro-Wilk test, we can consider that the model does comply with assumption 6.

Model 3 Interpretation

Covariates and r squared

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -8.3274644   4.3330321 -1.9219  0.059010 .
## prbarr       -1.7061418   0.3425156 -4.9812 4.941e-06 ***
## prbconv      -0.6740285   0.1396922 -4.8251 8.823e-06 ***
## prbpris      -0.0540833   0.4048599 -0.1336  0.894143
## avgsen       -0.0118876   0.0147992 -0.8033  0.424752
## polpc       223.7879810 122.3493915  1.8291  0.071974 .
## density       0.1082236   0.0573697  1.8864  0.063706 .
## taxpc         0.0013327   0.0067271  0.1981  0.843575
## west         -0.1611569   0.1220731 -1.3202  0.191410
## central      -0.1413052   0.0900223 -1.5697  0.121348
## urban        -0.1303464   0.2327918 -0.5599  0.577454
## pctmin80      0.0082900   0.0029275  2.8318  0.006155 **
## log(wcon)     0.1610118   0.2459475  0.6547  0.514997
## log(wtuc)     0.1639983   0.2862434  0.5729  0.568667
## log(wtrd)     0.1149604   0.3802354  0.3023  0.763358
## log(wfir)    -0.2059787   0.3823559 -0.5387  0.591928
## log(wser)    -0.4396133   0.2976624 -1.4769  0.144535
## log(wmfg)     0.0466298   0.1769835  0.2635  0.793022
## log(wfed)     0.9687255   0.4679283  2.0702  0.042407 *
## log(wsta)    -0.3602335   0.3637116 -0.9904  0.325633
## log(wloc)     0.3581009   0.6767133  0.5292  0.598485
## mix          -0.3909159   0.5597483 -0.6984  0.487432
## pctymle       1.9428112   2.1386436  0.9084  0.367006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "adjusted r-squared:  0.796622812092616"
```

```
## [1] "AIC =  16.2327644527931"
```

The first thing to note is adding coefficients has only slightly improved adjusted R^2 from 0.725 (Model 2) to 0.797. Additionally, we see that very few coefficients have a statistically significant impact. Considering that the model violates assumption 4, we do not consider this model to be robust.

Conclusion

The model comparrison table shows the three models side by side.

Table 2: Linear Models Predicting Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
prbconv	−0.392	−0.635	−0.674
prbpris			−0.054
avgsen		−0.003	−0.012
polpc			223.788
density	0.202	0.110	0.108
taxpc			0.001
west		−0.427	−0.161
central		−0.263	−0.141
urban			−0.130
pctmin80			0.008
log(wcon)			0.161
log(wtuc)			0.164
log(wtrd)			0.115
log(wfir)			−0.206
log(wser)			−0.440
log(wmfg)			0.047
log(wfed)		1.051	0.969
log(wsta)		−0.285	−0.360
log(wloc)		0.400	0.358
mix			−0.391
pctymle			1.943
prbarr		−1.548	−1.706
Constant	−3.609	−9.692	−8.327
AIC	83.772	32.927	16.233
Observations	88	88	88
R^2	0.484	0.753	0.848
Adjusted R^2	0.472	0.725	0.797
Residual Std. Error	0.379 (df = 85)	0.273 (df = 78)	0.235 (df = 65)

Model Robustness

From above comparision, we can conclude Model 2 is robust given:
1. Model 2 conforms to the six MLR assumptions while model 3 violates assumption 3 (Perfect Collinearity) and assumption 4 (Exogeneity).
2. AIC of Model 2 is efficient with an AIC of 32.9. 3. The signs of coefficient are same as Model 3. There is no change in signs for coefficients.
4. Model 2 holds a balance between parsimony and efficiency.

Even though Model 3 has a moderately improved r-squared, we think that Model 2 represents a better model for assessing crime rate in North Carolina, as it is more robust than Model 3 and more performant than Model 1 by comparing AIC.

Recommendations

Based on coefficients of Model 2, we conclude that density has positive relation on crime rate while probability of arrest, probability of conviction and average sentence have negative impacts on the crime rate. Dummy variables such as west and central also have negative impact on crime rate.

To reduce crime rate, we reccomend the following policies:

1. Focus on highly dense areas: Based on the correlation between density and crime rate, it is clear there may be a connection between these variables. Although it is not realistic to draft policy to change the density of a given location, there may be an opportunity to conduct further research into ommitted variables that are driving the positive correlation and effect density has on crime rate.

2. Focus on non-west and non-central areas: In the second model, we see the negative coefficients for central and west suggest counties that are included in that definition have a lesser crime rate. As the simple naming convention would not lead to reduced crime, there are ommitted variables that add to the strength of those negative effects. Since we do not know these variables, we suggest performing more research into omitted variables that may attribute for negative relationship between crime rate and west/central. While this research is being conducted, newly drafted policies could focus on counties not labeled as western or central, as these do not benefit from the negative effect on crime.

3. Appropriate more funding to train police: Based on Model 2, we see that the probability of conviction and the probability of arrest both have negative coefficients. As an increase in either of these variables will cause a decrease in crime rate, policy should be drafted in order to increase the ratio of arrests to offenses an increase the ratio of convictions to arrests. By giving police officers more training, precincts should see their officers know which arrests will lead to convictions and reduce the number of nessessary arrests.

4. Reconsider mandatory minimum sentences and scale of punishment policies: Model 2 shows the average sentence of a county has a negative effect on its crime rate, so increasing the severity of prison sentences and severity of crimes should result in a redution in crime rate. The team recommends lengthing prison sentences and increasing punishment to create this intended effect, and intuitively, it makes sense that this policy would deter rational criminals from committing crimes if the risk of penalties are more severe.