Multivariate Linear Regression on 1987 Crime Data in North Carolina

by

Patricia Wang, Tracy Stephens, Lester Yang

in the

Graduate Division

of the

University of California, Berkeley

datascience@berkeley

Contents

Contents **i**

1 Introduction **1**
1.1 About this Research 1
1.2 Initial Exploratory Data Analysis 1

2 Model Building: Base Model **9**
2.1 The Base Model 9
2.2 Regressing $ftfrte$ with the Base Model 14
2.3 Regressing *orte* with the Base Model 16
2.4 Comparing the Base Models 17

3 Model Building: Second Model **19**

4 Model Building: Third Model **25**

5 Conclusion **29**
5.1 Comparing Models 29
5.2 Omitted Variable Bias 31
5.3 Advised Policy 31

Chapter 1

Introduction

1.1 About this Research

This research aims to help the incumbent mayoral campaign in Raleigh, North Carolina to generate actionable policy related to the following research question:

What is the best way for a mayor to reduce crime rates in the city of Raleigh, North Carolina?

Raleigh, North Carolina, a mid-sized city with a population of just under 500,000, which has considerable problem with crime. According the FBI's crime Index, the city falls in the 11th percentile vs. the nation, meaning that 89 percent of US cities are safer than Raleigh. Therefore, our goal is to aid the campaign to propose policies that aim to decrease crime rates in Raleigh with insights gained by studying factors related to crime across the counties in the state. We intend to focus primarily on actionable policies.Therefore, our objective with this analysis is to identify variables that could reduce crime but, more importantly, over which a mayor have control. For example, if our candidate was running for sheriff, variables like the 'probability' of arrest would be of primary interest. Since our candidate is running for mayor, the set of controllable variables expands to include other variables, such as police per capita. A mayor might even have limited influence over other judiciary variables such as 'probability' of conviction, 'probability' of prison sentence, and average sentence length. Additionally, it is important to keep in mind that North Carolina is one of the most rural states in 1987 America, with urban population at 48% in 1980 and 57.8% in 1990 according to the U.S. Census Bureau. The suggested policies will account for this factor.

1.2 Initial Exploratory Data Analysis

We have been provided a cross-sectional data set and its corresponding codebook. This data set was first used in 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.) We aim for causal estimates of crime rates with ordinary least squares regression without looking further into the paper.

variable	label
1 county	county identifier
2 year	1987
3 crmrte	crimes committed per person
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
9 density	people per sq. mile
10 taxpc	tax revenue per capita
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
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
24 mix	offense mix: face-to-face/other
25 pctymle	percent young male

Table 1.1: Codebook of crime data set in North Carolina.

To measure crime rate, the data set enable us to operationalize crime rate as crimes committed per person with the field *crmrte*. Two categories of factors influence people to commit crime: environmental factors, and individual factors. In the following this study aims to regress *crmrte* on the environmental regressands: certainty and severity of punishment (proxied by the 'probability' variables, *polpc*, and *avgsen*), and living conditions (proxied by *density*, *taxpc*, area, and wage variables), and the individual regressands: income (proxied by wage variables), age/sex (proxied by *pctymle*), race/ethnicity (proxied by *pctmin*80) , and

the type of committed crime (proxied by mix). Based on this operationalization, this study expects to cover most of the explanatory factors of crime rate. Unfortunately, we are offered limited information on how the data was collected and how the codebook is annotated. The resulting limitation includes a lack of understanding regarding the data reporting rules on variables such as the certainty of punishment, and a skeptical interpretation of variable labels; al of which we will address in the following.

Initial examination of the data set revealed several issues in the data:

- 6 rows contained NAs on all columns
- *prbconv* has the type $factor$ instead of *double*.
- *prbarr*, and *prbconv* contain values above 1, deviating from mathematical definition of probability.
- *pctmin*80 contains values above 1, even though the similarly labeled *pctymle* contains only values below 1.
- County 185 has an extremely high $wser$.
- County 193 is duplicated.
- County 71 is labeled as both *central* and *west*.

To take care of this, we remove any rows with NA in the *county* column, as well as any duplicate values. Reviewing the "probability" fields ($prbarr, prbpris, prbconv$), as defined in the codebook, reveals that these fields are not defined as traditional probabilities bounded by $[0, 1]$. *prbconv* is the ratio of convictions to arrests, thus common sense suggests that it can be greater than 1 because a criminal can be convicted on multiple charges in a single arrest. *prbarr* on the other hand is not likely to be greater than 1 because that would suggest a criminal can be arrested multiple times for a single offense. However, we do not know how the data is collected. A scenario where *prbarr* is greater than 1 could result from a criminal escaping after their arrest has been recorded, but was arrested again for the same offense. Nonetheless, common sense would suggest that the two records of arrests should be one, and any occurrences otherwise should count as data entry error. Lastly, *prbpris* should be bounded by $[0, 1]$ because it is proxied by the convictions resulting in a prison sentence to total convictions.

With limited information on how the data labels are annotated, we make the assumption that the data set is labeled by the same person with consistency in mind. Since *pctmin*80 and *pctymle* are written in the same fashion, "pct" for percentage, as apposed to "prb", it is safe to assume that the inconsistency is simply caused by a data entry error. This study thus divide *pctmin*80 by 100 to match the expression method of percentage with *pctymle*.

The extreme $wser$ value is possibly a data entry error. Even though wages in the service sector should be reasonably consistent across counties, the distribution of wage is often right-tailed, and an extreme valid outlier can skew the distribution significantly. For example, it

is possible that a local restaurant tycoon lives in county 185. To account for this, the value is replaced by the mean of *wser*.

After initial cleaning (removing NA rows, duplicates), the data contained counties with 1 labeled as both *central* and *west*, 35 labeled as neither *central* nor *west*, 21 labeled as *west*, and 33 labeled as *central*. County 71 is removed because it is unclear which regions it belongs in, and with limited data rows (*west* < 30 data), relabeling can skew the results.

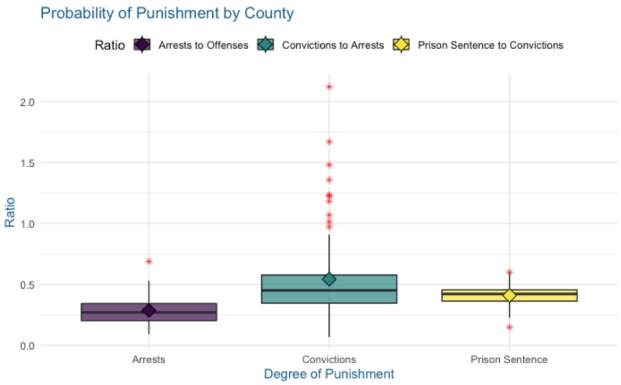
variables	Min	Mean	Max	SD	Relative SD	Max SD from Mean
crmrte	0.006	0.033	0.099	0.019	0.567	3.483
prbarr	0.093	0.296	1.091	0.138	0.468	5.747
prbconv	0.068	0.554	2.121	0.355	0.639	4.419
prbpris	0.150	0.411	0.600	0.081	0.197	3.220
avgsen	5.380	9.671	20.700	2.845	0.294	3.876
polpc	0.001	0.002	0.009	0.001	0.584	7.382
density	0.000	1.397	8.828	1.486	1.063	5.000
taxpc	25.693	38.235	119.761	13.167	0.344	6.192
pctmin80	0.000	0.003	0.006	0.002	0.659	2.261
wcon	193.643	285.285	436.767	48.019	0.168	3.155
wtuc	187.617	408.834	613.226	75.238	0.184	2.940
wtrd	154.209	210.591	354.676	33.916	0.161	4.248
wfir	170.940	321.325	509.466	54.231	0.169	3.469
wser	133.043	253.747	391.308	43.912	0.173	3.133
wmfg	157.410	335.775	646.850	88.697	0.264	3.507
wfed	326.100	441.868	597.950	59.864	0.135	2.607
wsta	258.330	357.725	499.590	43.539	0.122	3.258
wloc	239.170	311.973	388.090	28.140	0.090	2.705
mix	0.020	0.129	0.465	0.082	0.635	4.086
pctymle	0.062	0.084	0.249	0.024	0.280	6.983

Table 1.2: Descriptive statistic and standard deviation of all metric variables.

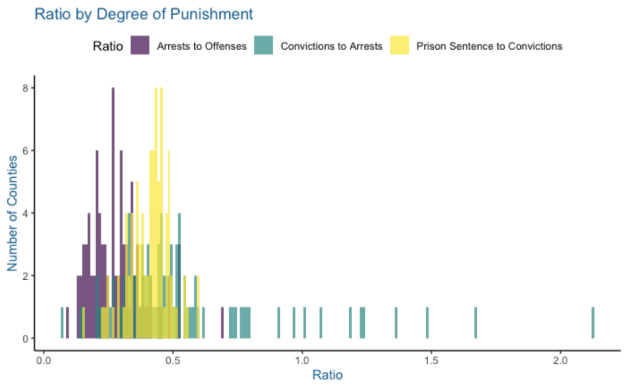
In Table 1.2, variables *prbbarr*, *polpc*, *density*, *taxpc*, and *pctymle* showed values that are 5 or more standard deviations away, representing the significant outliers. However, the number of data points are limited as explained, and we will thus examine the outliers when building models. To build the first model in accordance with the earlier section, let us begin with variables that policies have significant influence over: certainly and severity of punishment.

variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
prbarr	0.0928	0.2038	0.2715	0.2868	0.3439	0.6890
prbconv	0.0684	0.3466	0.4517	0.5437	0.5777	2.1212
prbpris	0.1500	0.3632	0.4222	0.4100	0.4560	0.6000
avgsen	5.3800	7.3500	9.0450	9.5456	11.3750	17.4100
polpc	0.0007	0.0012	0.0015	0.0016	0.0019	0.0045

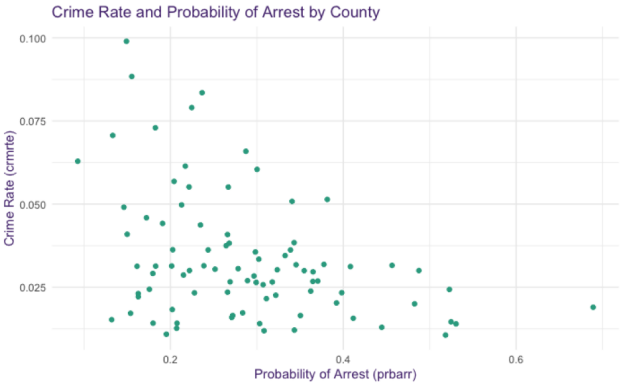
(a) Summary statistics of probability



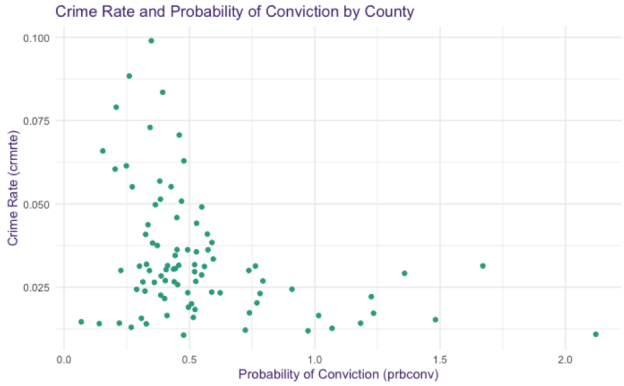
(b) Box Plot of 'Probability' Fields



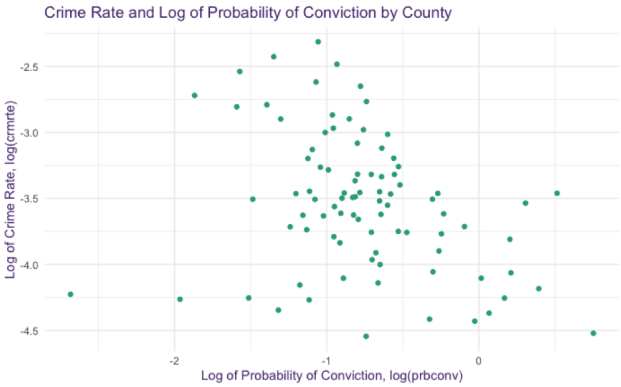
(c) Histogram of 'Probability' Fields



(d) Scatter Plot of prbarr

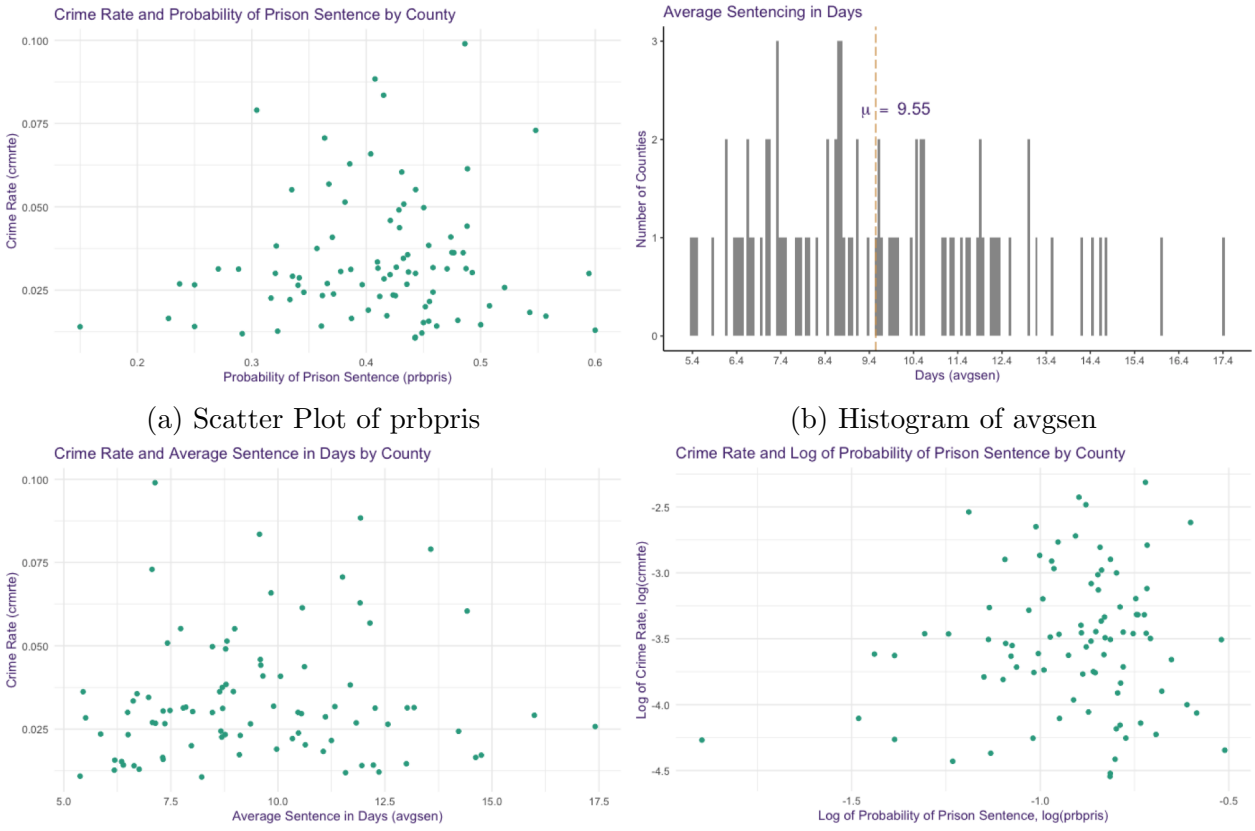


(e) Scatter Plot of log(prbconv)



(f) Scatter Plot of log(prbconv)

Figure 1.1: Exploratory Data Analysis of Certainty and Severity of Punishment



(a) Scatter Plot of prbpris

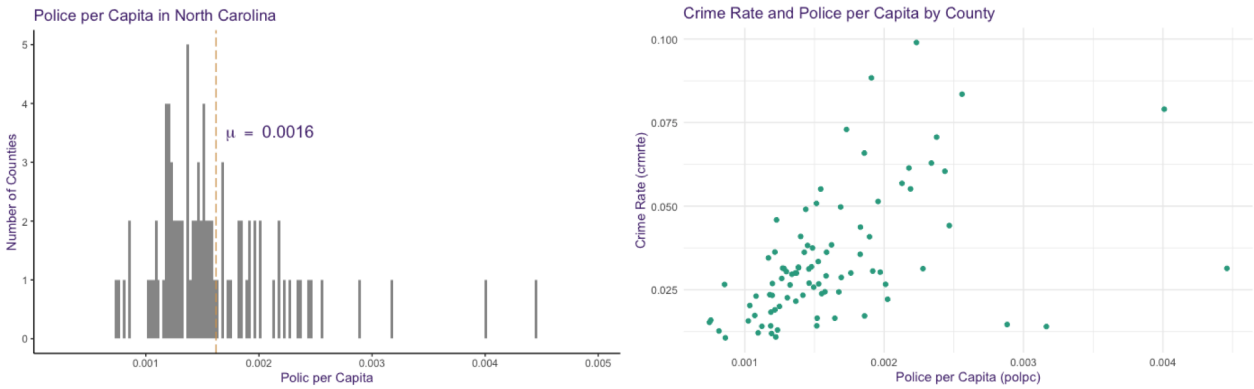
(b) Histogram of avgsen

(c) Scatter Plot of avgsen

(d) Scatter Plot of log(prbpris)

Figure 1.2: Exploratory Data Analysis of Certainty and Severity of Punishment (cont.)

Based on 1.1b and 1.1c the outliers of *prbconv* resulting from values near and greater than 1 are observed. From 1.1d and 1.1e, *prbarr* and *prbconv* appears to be negatively correlated with *crmrte*, observed as the decreasing in vertical variance with increasing "probability" in the scatter plot. Meanwhile, *prbpris* seems to be positively correlated with *crmrte* based on 1.2a. Figure 1.2b shows that *avgsen* has a small positive skew, while the exact correlation between *avgsen* and *crmrte* is not obvious based on the scatter plot (1.2c) as the it shows no obvious change in pattern above/below mean of 9.69 days. *polpc* is also related to the certainty of punishment as it is expected that increased police per capita increase the chance of criminals being arrested.



(a) Histogram of polpc

(b) Histogram of avgsen

Figure 1.3: Exploratory Data Analysis of Police per Capita

Excluding outliers, *polpc* appears to have a strong positive correlation with *crmrte*. One interpretation of this relationship is simply that more police were hired to deal with the higher crime rates of certain counties, a sort of a reverse causality. The few outliers further substantiate this interpretation because while there are counties with higher police per capita with lower crime rate, there are also counties with higher police per capita and higher crime rate. We expect this simultaneity bias contribute to the endogeneity of *polpc*, along with the related omitted variable bias.

For the dependent variable, *crmrte*, this study looks at crime rate individually, as well as combined with *mix* to show the distribution of 'face-to-face' versus other crimes. The *mix* field defines a ratio of crimes involving face-to-face contact to other crimes. How the crime rates of individual counties are regressed on the independent variables of interest should be affected by the mix field. As an example, it is possible that crime in high population density areas (high *density*) to have more face-to-face crimes, while more of other crimes occur in rural areas with lower population density (low *density*). Due to these possible factors, we view *mix* as an additional descriptor of crime rate, and will apply it to *crmrte* in order to view direct contact and other crime rates separately. Specifically, new variables $ftfrte$ and *orte* are created for 'face-to-face' crimes and other crimes, respectively, with the following equations:

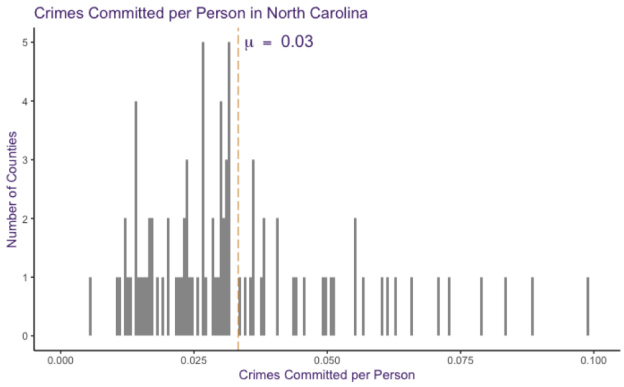
$$mix = \frac{\text{face-to-face crimes}}{\text{other crimes}} \Rightarrow mix_{prob} = \frac{mix}{1 + mix}$$

$$ftfrte = mix_{prob} \times crmrte$$

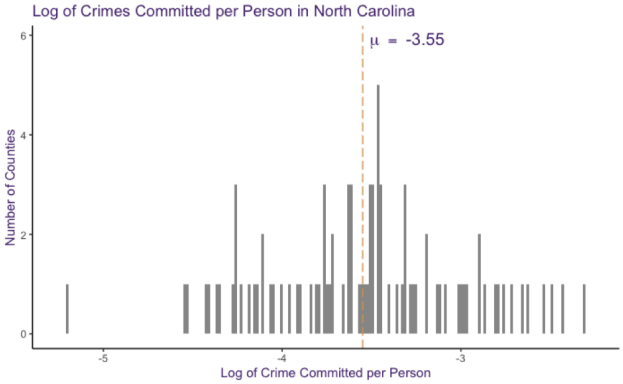
$$orte = (1 - mix_{prob}) \times crmrte$$

variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
crmrte	0.0055	0.0203	0.0300	0.0333	0.0384	0.0990
ftfrte	0.0003	0.0019	0.0032	0.0036	0.0049	0.0143
orte	0.0088	0.0180	0.0261	0.0300	0.0350	0.0847

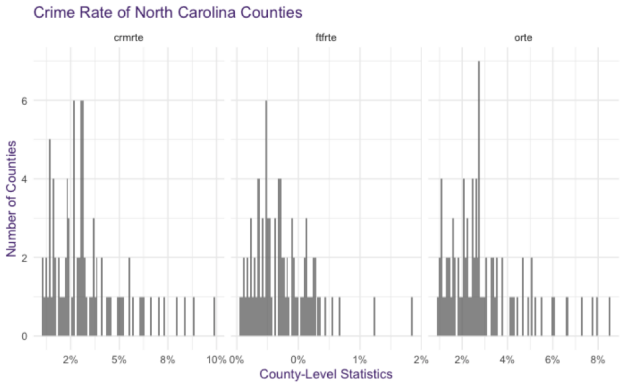
(a) Summary statistics of crmrte



(b) Histogram of crmrte



(c) Histogram of Log Transformed crmrte



(d) Histograms of mix-Applied crmrte

Figure 1.4: Exploratory Data Analysis of Crime Rate

Notice the rates of non-violent crimes are much higher than the rates of violent crimes. All three measures are positively skewed, as would be expected with proportions, which are bounded at 0. It is notable that theoretically, crime rates could be above 1, since a person can commit more than 1 crime per year. However, this is almost never the case when we look at the *crmrte* values. To transform the positively-skewed distribution of the crime rate variables, we take their natural log. The new distribution, as observed in Figure 1.4c, is less skewed.

Chapter 2

Model Building: Base Model

2.1 The Base Model

For the base model, we only include variables for which we believe a mayor would be able to exercise change, which generally, includes the variables that affect the certainty and severity of punishment:

$$\begin{aligned} log(crmrte) = \beta_0 + \beta_1\ log(prbarr) + \beta_2\ log(prbconv) + \beta_3\ log(prbpris) \\ + \beta_4\ avgsen + \beta_5\ log(polpc) + u \end{aligned}$$

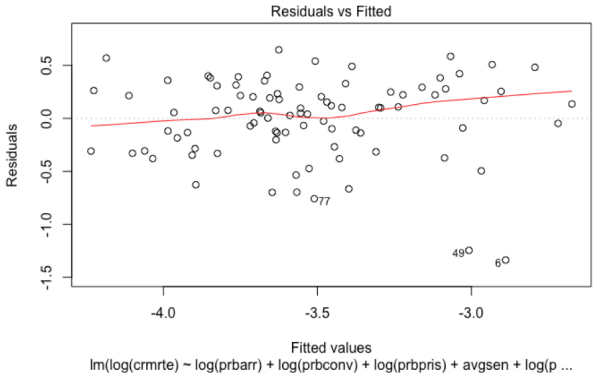
for which we here regress $log(crmrte)$ on $\{log(prbarr), log(prbconv), log(prbpris), avgsen, log(polpc)\}$ for the slope estimators. Since *crmrte* has been transformed to $log(crmrte)$ based on its skewness, *prbarr*, *prbconv*, *prbpris*, and *polpc* are also transformed to the logarithmic form for ease of interpretation. In this log-log form, the regression coefficients β are elasticities, and are interpreted as 1% increase in regressors $(prbarr, prbconv, polpc)$ would on average lead to a *ceteris paribus* β% increase in the regressand $(crmrte)$. An additional reason for log-transforming *prbconv* is observed in Figure 1.1e as the near exponentially decreasing relationship, and the positively skewed distribution in Figure 1.1c. The log-transformed *prbconv* gave the scatter plot in Figure 1.1f. In addition to linearizing the relationship with log-transformed *crmrte*, noticeable outliers were created. They are expected to influence the model fit.

Mean and Median values in Table 1.2 suggests that among the punishment variables, *prbarr*, *prbconv*, and *avgsen* are positively skewed. *prbpris* has a minor negative skew, and *polpc* is essentially not skewed. *avgsen*, despite its larger skew compared to *prbarr*, and even *polpc*, was not log-transformed because often times skewness alone is not enough evidence to justify transformation. Interpretation is also important. For example, *polpc* was transformed in this case because a percentage change in police per capita is more interpretable. On the other hand, keeping *avgsen* as the regressor means that for its estimate β, a 1 percent change in crime rate while holding other independent variables constant has a 100β effect on the average sentence in days. This non-percentage change interpretation is more intuitive,

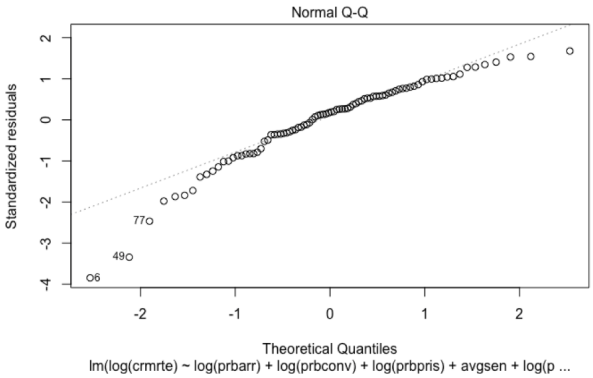
and thus preferred. The coefficient estimates and diagnostic plots of this base model are generated below in Figure 2.1.

(Intercept)	log(prbarr)	log(prbconv)
0.0551	-0.5520	-0.3118
log(prbpris)	avgsen	log(polpc)
0.2279	-0.0137	0.6498

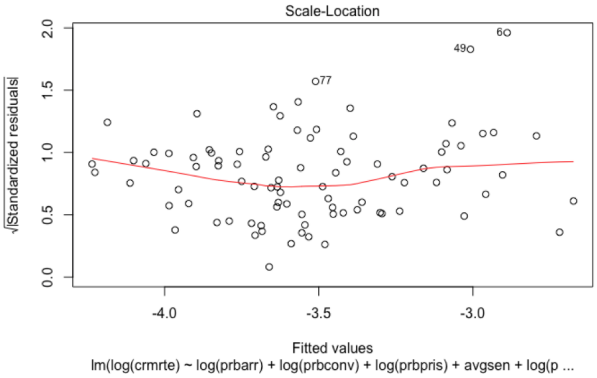
(a) Coefficient estimates



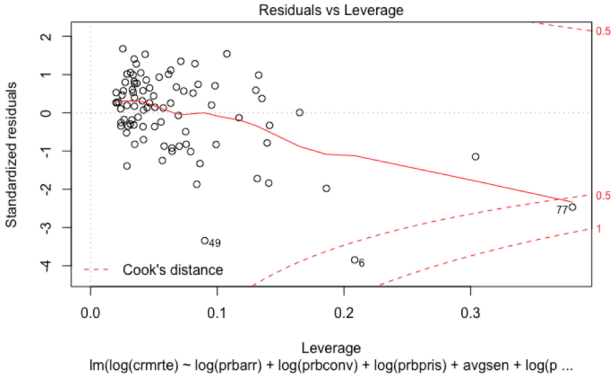
(b) Residuals vs. Fitted Values Plot



(c) Normal Q-Q Plot



(d) Scale-Location Plot



(e) Residual vs. Leverage Plot

Figure 2.1: Coefficient Estimates and Diagnostic Plots of the Base Model (*crmrte* regressed)

We first examine if this base model follow the 6 CLM assumptions.

1. MLR.1 Linear population model: We have yet to constrain the error term to be normal so checking is not required at this point.

2. MLR.2 Random sampling: Unfortunately, we are not provided how the data collected in the documentation. While individual fields for a county should be collected from different data sources, we assume that counties are sampled at random for generating the data set.

3. MLR.3 No perfect collinearity: *lm* function in R did not output alerts, and as such it is safe to assume the model satisfy this assumption.

4. MLR.4 Zero conditional mean: In Figure 2.1b, we can notice the red line segment generated by the mean of residual values shows that our coefficients are near unbiased. There is a small positive skew in counties with higher *crmrte* (higher $log(crmrte)$) possibly due to an omitted variable in the regression. For example, a possible variable in the error term u that correlates with *avgsen* is the average age of the convicted criminals in the county. In counties with a larger underage population or young adults involved in crimes, *avgsen* can be skewed. The younger population largely involved in petty crimes or misdemeanor can receive shorter sentences compared to felonies, and as such *avgsen* would be lower for such a county. Another possible variable is the average number of civilian owned firearms involved in crimes per county. This variable will be correlated with both *avgsen* and $(prbarr, prbconv, prbpris)$ because if more civilian owned firearms are involved in crimes, *avgsen* and the "probability" variables should all rise dramatically. These two omitted variables both cause an functional dependence between the independent variables and the error term. Further evaluation of omitted variable bias will be discussed in a later section.

5. MLR.5 Homoskedasticity: Figures 2.1b and 2.1d does not seem to indicate heteroskedasticity, except for a few outliers that have lower residuals, the model appears to be homoskedastic. Despite this we will use heteroskedasticity-robust errors in the following inference as it is good practice.

6. MLR.6 Normality of errors: Figure 2.1c and the following histogram shows that our residuals have a leftward skew. Although the Shapiro-Wilk's test produced a p-value of 0.0005263, suggesting that our distribution is non-normal, we can apply CLT based on our sufficient sample size (88 rows). As a result, as the sample size grows, our model should have a normal error distribution.

Diagnostic plots in 2.1 shows that observations 6, 49, and 77 might be potential problems. The Cook's distance measured in Figure 2.1e helps in understanding whether these observations are influential in determining the regression line. Specifically, observations 6 and 77 have moderate Cook's distance above 0.5. Comparing the outlier variable values with the rest of the observations in Tables **??**, we noticed that observations 6 and 77 contain the

minimum values of $log(prbconv)$, and $log(prbpris)$, respectively. These extreme values most likely contributes to their relatively high influence on this model. Despite this fact, we have decided not to remove observations 6 and 77, because the Cook's distance is not above 1, and both of these counties have $west = 1$. Since the number of $west = 1$ observations is already small (= 20), we have decided that removing these moderately influencing observations is not worth the trade off, as we are also expecting representative outliers, as explained in the introduction.

variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
log(prbarr)	-2.3776	-1.5905	-1.3039	-1.3189	-1.0675	-0.3725
log(prbconv)	-2.6827	-1.0596	-0.7947	-0.7606	-0.5488	0.7520
log(prbpris)	-1.8971	-1.0129	-0.8622	-0.9141	-0.7853	-0.5108
avgsen	5.3800	7.3500	9.0450	9.5456	11.3750	17.4100
log(polpc)	-7.2009	-6.6987	-6.5145	-6.4816	-6.2884	-5.4128

(a) Summary statistics of variables in the base model.

obs.	crmrte	log(prbarr)	log(prbconv)	log(prbpris)	avgsen	log(polpc)
6	0.0146067	-0.6449972	-2.682732	-0.6931472	13.00	-5.849260
77	0.0139937	-0.6340578	-1.115141	-1.8971199	6.64	-5.755985

(b) Variable values of influential outliers

Figure 2.2: Comparison of outlier variable values with the rest of the observations

We now assess to what extent multicollinearity is affecting our inference. Based on the variable inflation factor (VIF) calculated, there are little to no multicollinearity among the independent variables in this model.

log(prbarr)	log(prbconv)	log(prbpris)	avgsen	log(polpc)
1.2239	1.2908	1.0409	1.1255	1.2872

Table 2.1: Variable inflation factor of the base model.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.055114	1.607563	0.0343	0.972734
log(prbarr)**	-0.551985	0.166346	-3.3183	0.001352
log(prbconv) .	-0.311829	0.161498	-1.9309	0.056957
log(prbpris)	0.227926	0.359286	0.6344	0.527595
avgsen	-0.013683	0.023345	-0.5861	0.559389
log(polpc)**	0.649756	0.216313	3.0038	0.003535

Table 2.2: t-test of coefficients for the base model. `Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .`

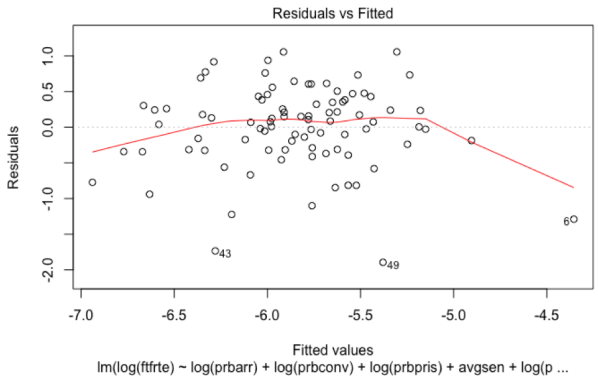
Based on the assumptions MLR. 1 to MLR. 5, our OLS estimators are the best linear unbiased estimators (BLUEs). Additionally with MLR. 6, we have built a classical linear model based on the classical linear model (CLM) assumptions. The t-test of estimates above show that $log(prbarr)$ and $log(prbconv$ are statistically significant at the 0.1% level, and $log(polpc)$ is statistically significant at the 5% level. Practically the model suggests that a 1% increase in either three of the statistically significant independent variables will lead to a *ceteris paribus* 0.552% decrease, 0.312% decrease, and 0.650% increase in *crmrte*, respectively. These changes are practically small. Additionally, the interpretation for *polpc* needs to account for the simultineity bias as previously mentioned. The causal effect of *polpc* on *crmrte* should be lower.

2.2 Regressing $ftfrte$ with the Base Model

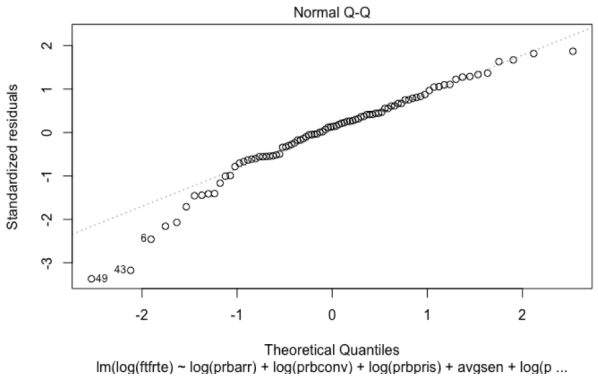
The coefficient estimates and diagnostic plots of the base model with *crmrte* replaced with $ftfrte$ are generated below in Figure 2.3.

(Intercept)	log(prbarr)	log(prbconv)
-0.0197	0.1067	-0.4862
log(prbpris)	avgsen	log(polpc)
0.3532	-0.0338	0.8354

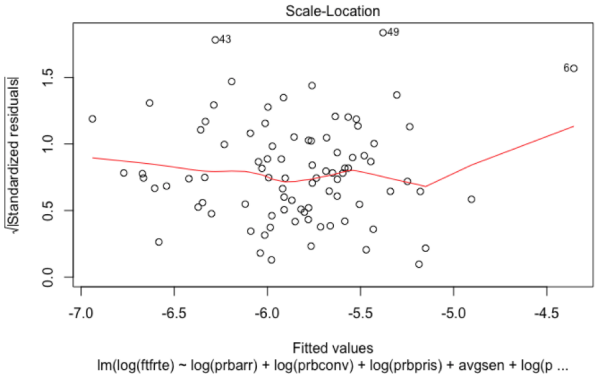
(a) Coefficient estimates



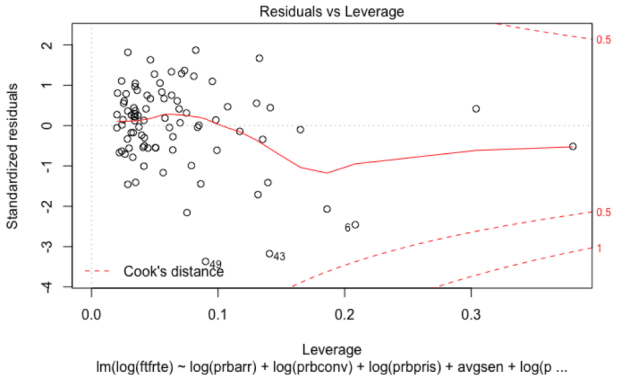
(b) Residuals vs. Fitted Values Plot



(c) Normal Q-Q Plot



(d) Scale-Location Plot



(e) Residual vs. Leverage Plot

Figure 2.3: Coefficient Estimates and Diagnostic Plots of the Base Model ($ftfrte$ regressed.)

Based on the same reasons for the base model above, this model satisfy MLR.1 to MLR.6 as well. Although the residual vs. fitted plot indicates with the red line segment that the residuals of this model deviates from zero mean for larger fitted values, the deviation is not significant, and it might be an artifact of less data points in the region as well. Also the Breusch-Pagan test returned a p-value of 0.2325, and we fail to reject the null hypothesis that assumes homoskedasticity. Also notice that for face-to-face crime rates, the observations no longer have highly influential points.

t test of coefficients $ftfrte$				
	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-0.019657	1.949366	-0.0101	0.99198
log(prbarr)	0.106717	0.230026	0.4639	0.64392
log(prbconv)*	-0.486197	0.214530	-2.2663	0.02606
log(prbpris)	0.353154	0.364919	0.9678	0.33601
avgsen	-0.033795	0.032499	-1.0399	0.30145
log(polpc)**	0.835404	0.245643	3.4009	0.00104

Table 2.3: t-test of coefficients for the base model regressing $ftfrte$. `Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .`

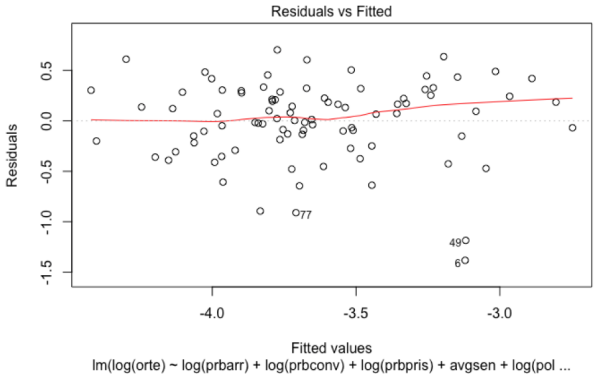
The coefficient t-test shows that $log(prbconv)$ is statistically significant at the level of 1%, and $log(polpc)$ statistically significant at the 0.1% level. Practically, a 1% increase in either $log(prbconv)$ or $log(polpc)$ leads to a small 0.488% decrease, or a small 0.836% increase in the log-transformed $crmrte$.

2.3 Regressing *orte* with the Base Model

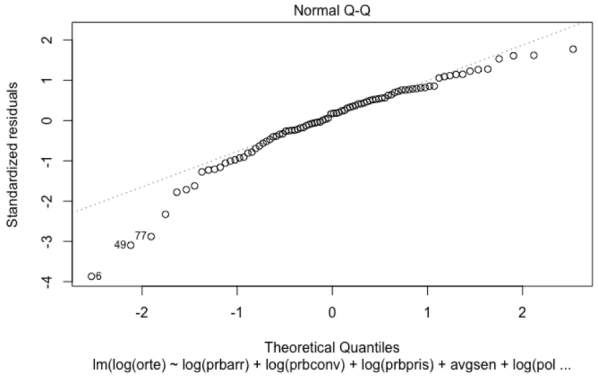
The coefficient estimates and diagnostic plots of the base model with *crmrte* replaced with *orte* are generated below in Figure 2.4.

(Intercept)	log(prbarr)	log(prbconv)
-0.4009	-0.642	-0.2915
log(prbpris)	avgsen	log(polpc)
0.2084	-0.0108	0.6208

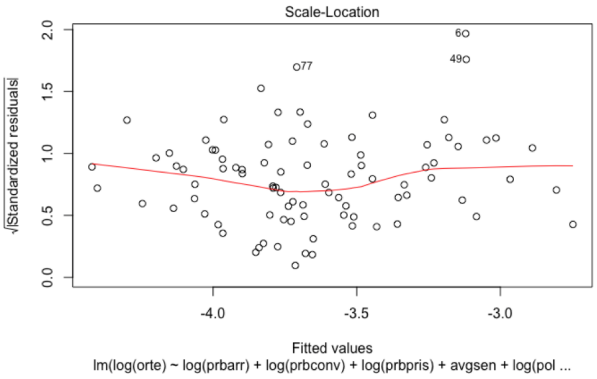
(a) Coefficient estimates



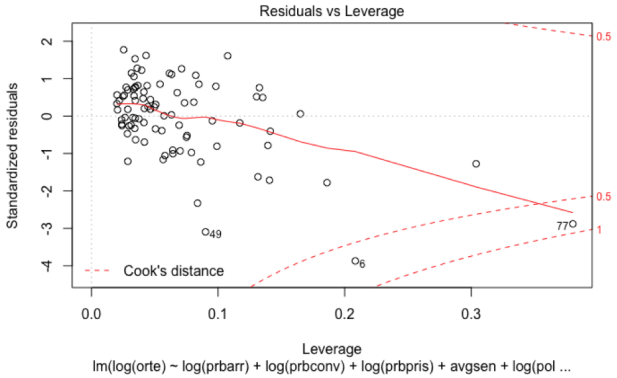
(b) Residuals vs. Fitted Values Plot



(c) Normal Q-Q Plot



(d) Scale-Location Plot



(e) Residual vs. Leverage Plot

Figure 2.4: Coefficient Estimates and Diagnostic Plots of the Base Model (*orte* regressed.)

Based on the same reasons for the base model above, this model satisfy MLR.1 to MLR.6 as well. We observe the same skew as observed in the base model on the normal Q-Q plot. And we can see that it satisfy both zero conditional mean and homoskedasticity from the residual vs. fitted plot. The moderately influential observations 6, and 77 was not removed for the same reason stated in the base model on *crmrte*.

t test of coefficients *orte*				
	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-0.400926	1.727845	-0.2320	0.8170864
log(prbarr)***	-0.642020	0.173499	-3.7004	0.0003884
log(prbconv) .	-0.291488	0.163412	-1.7838	0.0781620
log(prbpris)	0.208440	0.406261	0.5131	0.6092818
avgsen	-0.010787	0.024857	-0.4340	0.6654617
log(polpc)*	0.620781	0.237730	2.6113	0.0107237

Table 2.4: t-test of coefficients for the base model regressing *orte*. `Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1`

In this case regression on *orte* closely resembles that of *crmrte* in terms of independent variables that are statistically significant because *orte* contains a larger percentage of *crmrte* compared to $ftfrte$. Notice here that by removing the face-to-face crime contribution, $log(prbarr)$ is now highly significant compared to regression on *crmrte*.

2.4 Comparing the Base Models

Finally we examine all three variations of the base model and their corresponding fit with Table 2.5. What is immediately noticeable is that the same percentage increase in police per capita for face-to-face and other crimes corresponds to a higher percentage increase in face-to-face crime rates than other crime rates. This seems logical because face-to-face crimes are often times more violent and can thus lead to more police involvement. Readers should keep simultineity bias in mind when comparing the *polpc* estimates, all estimates should be above true values. Another statistically significant result across all three models is $log(prbconv)$, where we can see the same percentage increase in conviction rate leads to higher percentage decrease in face-to-face crime rate than other/overall crime rate. Meanwhile, the same percentage increase in arrest rate leads to a higher percentage decrease in other crime rate than the overall crime rate. In terms of model fit, since the dependent variables are different, comparing the model fit with R-squared would be meaningless. We reserve the comparison between models for the concluding chapter.

Comparing these three versions of the model, the differences in the coefficients for $log(prbarr)$ and $log(prbconv)$ might suggest that increasing rates of arrest decreases non-violent crimes more than violent crimes, while increasing rates of conviction decreases violent

	Dependent variable:		
	Overall Crime Rate	Violent Crime Rate	Non-Violent Crime Rate
	(1)	(2)	(3)
log(prbarr)	-0.552^{***}	0.107	-0.642^{***}
	(0.166)	(0.230)	(0.173)
log(prbconv)	-0.312^{*}	-0.486^{**}	-0.291^{*}
	(0.161)	(0.215)	(0.163)
log(prbpris)	0.228	0.353	0.208
	(0.359)	(0.365)	(0.406)
avgsen	−0.014	−0.034	−0.011
	(0.023)	(0.032)	(0.025)
log(polpc)	0.650^{***}	0.835^{***}	0.621^{***}
	(0.216)	(0.246)	(0.238)
Constant	0.055	−0.020	−0.401
	(1.608)	(1.949)	(1.728)
Observations	88	88	88
R^2	0.471	0.367	0.468
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 2.5: Linear models of crime rates with variations of the base model.

crimes more than non-violent crimes. Therefore, we would point out to the mayor that since decreasing violent crimes may be more politically advantageous, it may make more sense to focus resources on increasing conviction rates than arrest rates alone.

Chapter 3

Model Building: Second Model

In this model, we add in key demographic variables that likely correlate highly with crime rates, but that a mayor wouldn't necessarily have control over. These include *density*, *pctmin*80, *pctymle*, and *wages*.

As with the base model, we choose to transform the independent variables, *density*, *pctmin*80, and *pctymle*, for ease of interpretation, making the regression coefficients β elasticities.

Then, for the wage variables, we performed some analysis to identify a good way to combine them.

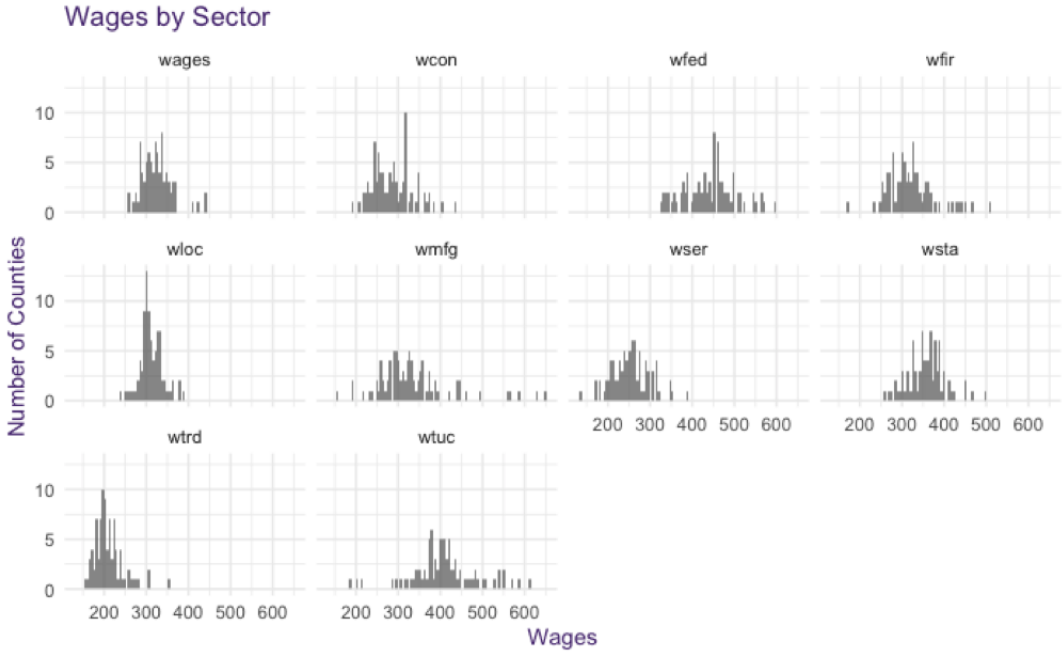


Figure 3.1: Histogram of Wages by Sector

All of the wage variables are approximately normally distributed, with no significany outliers (since we accounted for outlier of *wser* earlier).

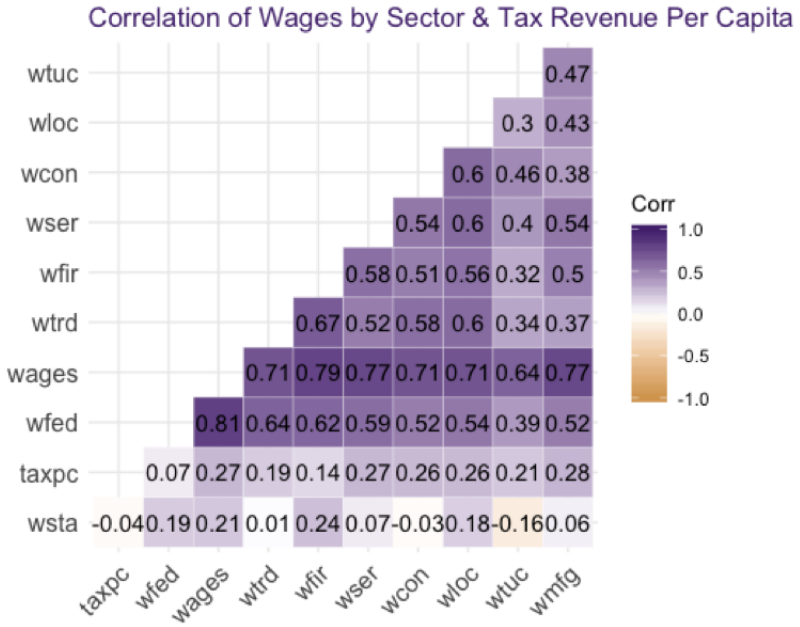


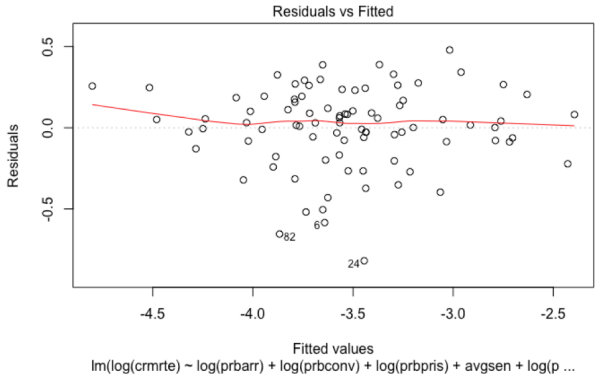
Figure 3.2: Correlation Heatmap of Wages

Additionally, nearly all the wage variables, with the exception of state employees (*wsta*) are highly correlated. The state government likely sets wages of state employees, so they are less likely to reflect economic conditions in individual counties than the other wage variables. To simplify the wage variables, we take an average of all of them. This assumes that the same number of people work in each sector, which of course is not true, but since all the sectors are correlated, this number should serve as a decent proxy for the true average weekly wage per capita in the county.

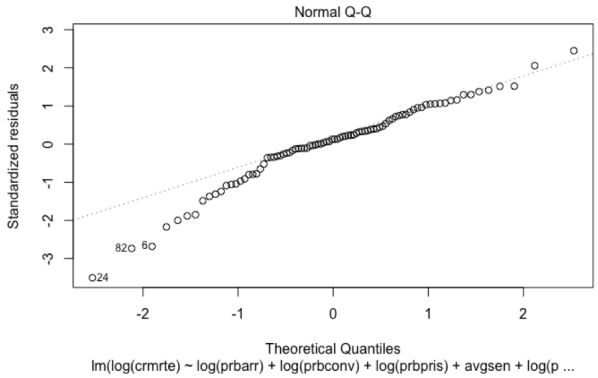
$$\begin{aligned} log(crmrte) =& \beta_0 + \beta_1\ log(prbarr) + \beta_2\ log(prbconv) + \beta_3\ log(prbpris) \\ &+ \beta_4\ avgsen + \beta_5\ log(polpc) + \beta_6\ log(density) + \beta_7\ log(wages) \\ &+ \beta_8\ log(pctmin80) + \beta_9\ log(pctymle) + u \end{aligned}$$

Intercept	log(prbarr)	log(prbconv)	log(prbpris)	avgsen
-2.9948	-0.4091	-0.3414	-0.2550	-0.0235
log(polpc)	log(density)	log(wages)	log(pctmin80)	log(pctymle)
0.5722	0.1422	0.4894	0.2411	0.0198

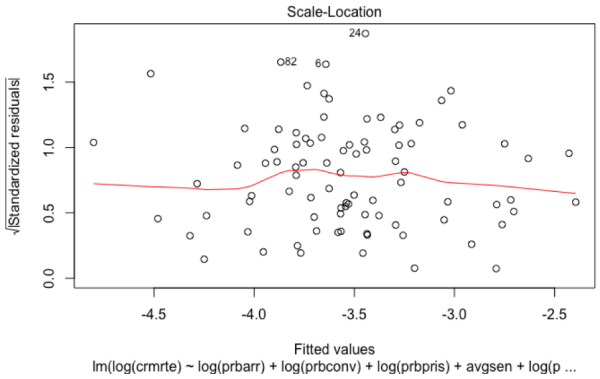
Table 3.1: Coefficient estimates for the second model.



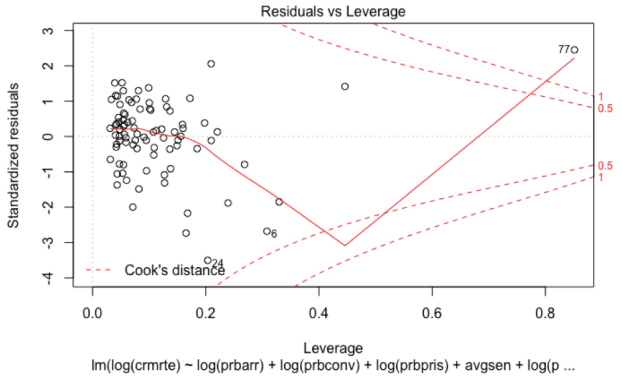
(a) Residuals vs. Fitted Values Plot



(b) Normal Q-Q Plot



(c) Scale-Location Plot



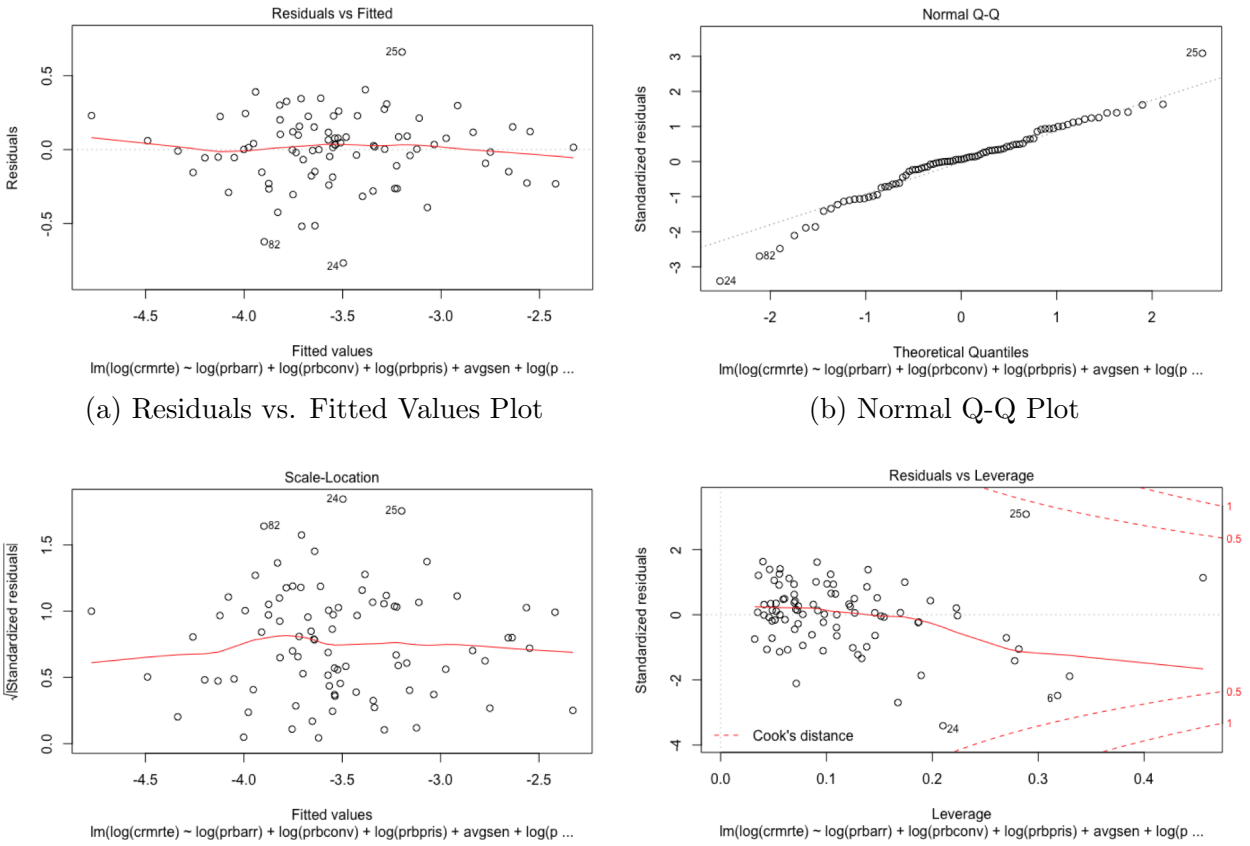
(d) Residual vs. Leverage Plot

Figure 3.3: Coefficient Estimates and Diagnostic Plots of the Second Model

The Cook's distance for observation 77 in shown in Figure 3.3d in now above 1, too large a value to justify keeping the observation for this model. Observation 77 is thus removed for remodeling as it also showed up as an outlier in Figure 1.2.

Intercept	log(prbarr)	log(prbconv)	log(prbpris)	avgsen
-0.9504	-0.3731	-0.3101	-0.1781	-0.0178
log(polpc)	log(density)	log(wages)	log(pctmin80)	log(pctymle)
0.4867	0.2689	0.0411	0.2373	-0.0064

Table 3.2: Coefficient estimates for the second model.



(a) Residuals vs. Fitted Values Plot

(b) Normal Q-Q Plot

(c) Scale-Location Plot

(d) Residual vs. Leverage Plot

Figure 3.4: Coefficient Estimates and Diagnostic Plots of the Second Model, removing obs. 77.

With the new coefficients and diagnostic plots in Figure 3.4, we examine if this second model follows the 6 CLM assumptions.

1. MLR.1 Linear population model: We have yet to constrain the error term to be normal so checking is not required at this point.

2. MLR.2 Random sampling: Unfortunately, we are not provided how the data collected in the documentation. While individual fields for a county should be collected from

different data sources, we assume that counties are sampled at random for generating the data set.

3. MLR.3 No perfect collinearity: *lm* function in R did not output alerts, and as such it is safe to assume the model satisfy this assumption.

4. MLR.4. Zero conditional mean: In Figure 3.3a, we can notice the red line segment generated by the mean of residual values show that our coefficients are near unbiased.

5. MLR.5. Homoskedasticity: Figures 3.3a and 3.3c does not seem to indicate heteroskedasticity. Despite this we will use heteroskedasticity-robust errors in the following inference as it is good practice.

6. MLR.6 Normality of errors: Figures 3.3b shows that our residuals have a slight rightward skew. Applying the Shapiro-Wilk's test produced a p-value of 0.08207, we fail to reject the null hypothesis at the 5% level, and assumes normality.

The variable inflation factor of the second model shows that the increased number of explanatory variables began to incorporate some multicollinearity among the variables. In particular, $log(density)$ and $log(wages)$ have VIF values above 2.

log(prbarr)	log(prbconv)	log(prbpris)	avgsen	log(polpc)
1.603063	1.579686	1.099304	1.224044	1.698734
log(density)	wages	log(pctmin80)	log(pctymle)	
2.541134	2.335356	1.059550	1.293557	

Table 3.3: Variable inflation factor of the second model.

t test of coefficients				
	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	-0.950436	3.721870	-0.2554	0.7991219
log(prbarr) **	-0.373060	0.118676	-3.1435	0.0023711
log(prbconv) *	-0.310082	0.122402	-2.5333	0.0133313
log(prbpris)	-0.178075	0.199599	-0.8922	0.3750854
avgsen	-0.017786	0.012245	-1.4526	0.1503994
log(polpc)*	0.486665	0.227778	2.1366	0.0358087
log(density) ***	0.268863	0.077457	3.4711	0.0008529
log(wages)	0.041138	0.510654	0.0806	0.9360011
log(pctmin80) ***	0.237287	0.038773	6.1199	3.647e-08
log(pctymle)	-0.006449	0.195337	-0.0330	0.9737482

Table 3.4: t-test of coefficients for the second model. `Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .`

Model 2 indicates that *prbarr*, *prbconv*, and *polpc* are still significant, but *density* and *pctmin*80 are statistically significant as well. *pctmin*80 has a positive regression coefficient, which indicates that counties with higher minority populations tend to also have higher crime rates. More concretely, with a 1% increase in pctmin80, the model indicates a 0.24% increase in crime rate. Similarly *density* is highly significant, with a 1% increase corresponding to a 0.27% increase in crime rate. In terms of interpretation, adding the demographic variables slightly lessens the impact of increasing rates of arrest on crime rates, but increases the estimated impact of increasing the probability of conviction.

There is some shift in the coefficients from Model 1 to Model 2, most notably with *prbarr* and *polpc*. So, for a 1% increase in:

- *prbarr*, we see a shift from 0.55% (Model 1) to 0.37% (Model 2) in decrease in crime rate
- *polpc*, we see a shift from 0.65% (Model 1) to 0.49% (Model 2) in increase in crime rate

It is also worth noting that there is a notable increase in the $adjustedR^2$ from Model 1 at 0.438 to Model 2 at 0.761. That said, since we added more variables to Model 2, we also examined how Model 1 and Model 2's AIC values compared. Model 1's AIC is at 92.1 and Model 2's AIC is at 19.5. Since Model 2's AIC is significantly lower than that of Model 1, we conclude that Model 2 is the more parsimonious model between the two.

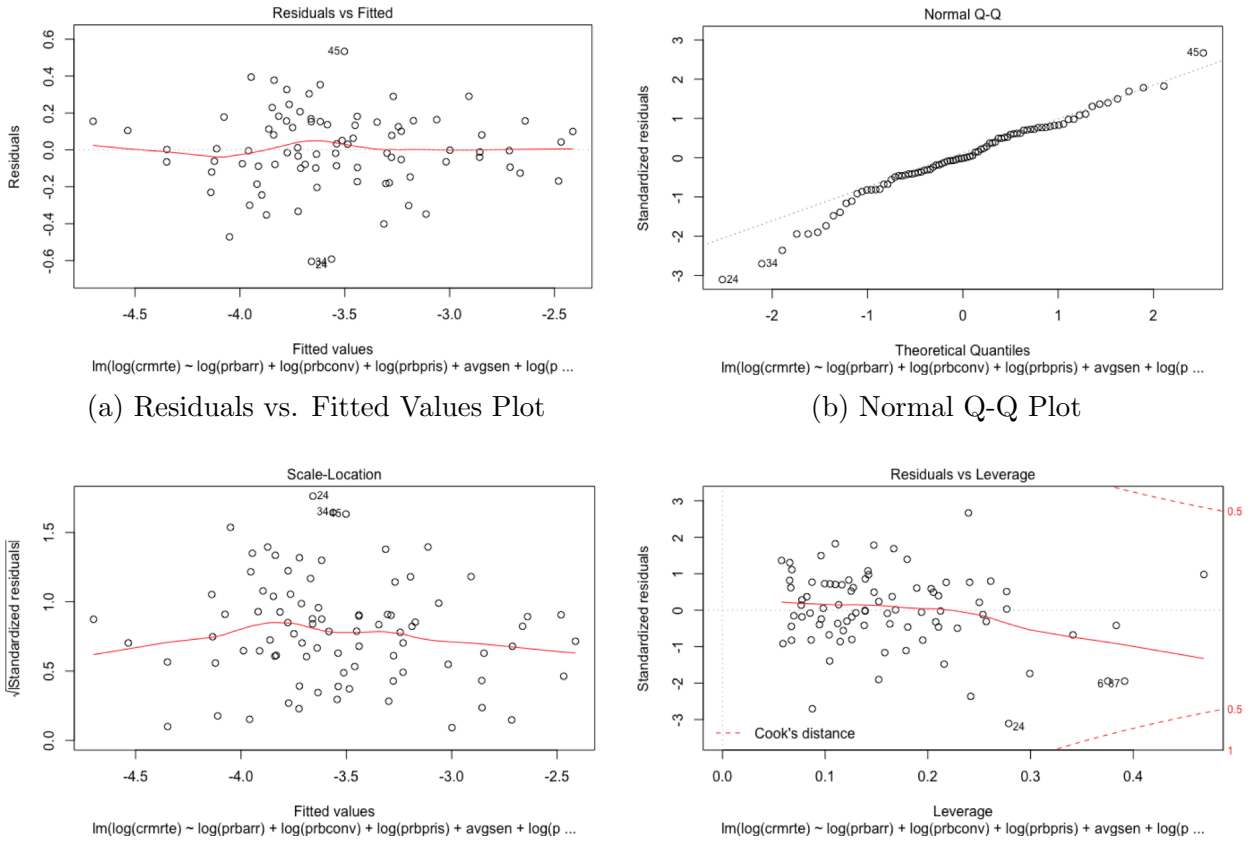
Chapter 4

Model Building: Third Model

For the third model, we add in all the variables, which includes *taxpc* and the location variables *west*, *central*, and *urban*, in addition to the variables already included. The initial diagnostic plots showed again that observation 77 have Cook's distance above 1, and based on the reasons previously mentioned, we removed observation 77. In the obs. 77 excluded model observation 25 had Cook's distance above 1. Examining obs. 25 showed that it should also be removed since it is not a Western, Central, or urban county. We argue that we have more observations that are neither Western nor Central, and since it is not a urban county as well, it is less likely to be a representing outlier in the mostly rural North Carolina.

Intercept	log(prbarr)	log(prbconv)	log(prbpris)	avgsen
-3.1325	-0.3436	-0.2950	-0.0671	-0.0235
log(polpc)	log(density)	wages	log(pctmin80)	log(pctymle)
0.3728	0.3642	0.3617	0.2587	-0.0393
taxpc	west	central	urban	
-0.0051	-0.0136	-0.1661	-0.0696	

Table 4.1: t-test of coefficients for the second model. `Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .`



(a) Residuals vs. Fitted Values Plot

(b) Normal Q-Q Plot

(c) Scale-Location Plot

(d) Residual vs. Leverage Plot

Figure 4.1: Coefficient Estimates and Diagnostic Plots of the Third Model

Based on Figure 4.1, we re-examine if this third model follows the 6 CLM assumptions.

1. MLR.1 Linear population model: We have yet to constrain the error term to be normal so checking is not required at this point.

2. MLR.2 Random sampling: Unfortunately, we are not provided how the data collected in the documentation. While individual fields for a county should be collected from different data sources, we assume that counties are sampled at random for generating the data set.

3. MLR.3 No perfect collinearity: *lm* function in R did not output alerts, and as such it is safe to assume the model satisfy this assumption.

4. MLR.4 Zero conditional mean: In Figure 4.1a, we can notice the red line segment generated by the mean of residual values show that our coefficients are near unbiased.

5. MLR.5. Homoskedasticity: Figures 4.1a and 4.1c does not seem to indicate heteroskedasticity. Despite this we will use heteroskedasticity-robust errors in the following inference as it is good practice.

6. MLR.6 Normality of errors: Figures 4.1b shows that our residuals have a slight leftward skew. Applying the Shapiro-Wilk's test produced a p-value of 0.08207, we fail to reject the null hypothesis at the 5% level, and assumes normality.

t test of coefficients				
	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	-3.1324647	3.5272654	-0.8881	0.3774594
log(prbarr) **	-0.3436291	0.1026932	-3.3462	0.0013046
log(prbconv) **	-0.2949802	0.1059982	-2.7829	0.0068759
log(prbpris)	-0.0670723	0.1591517	-0.4214	0.6746923
avgsen .	-0.0235484	0.0121058	-1.9452	0.0556544
log(polpc) .	0.3727970	0.2005125	1.8592	0.0670820
log(density) ***	0.3642488	0.0826897	4.4050	3.613e-05
log(wages)	0.3616913	0.5213104	0.6938	0.4900327
log(pctmin80) ***	0.2586686	0.0659981	3.9193	0.0002002
log(pctymle)	-0.0392668	0.1521521	-0.2581	0.7970844
taxpc	-0.0051033	0.0055220	-0.9242	0.3584801
west	-0.0136217	0.1469548	-0.0927	0.9264047
central .	-0.1660611	0.0906377	-1.8321	0.0710660
urban	-0.0696097	0.1607697	-0.4330	0.6663251

Table 4.2: t-test of coefficients for the third model. `Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .`

In Model 3, only *central* in the extra added variables that were not in Model 2 proved to be statistically significant. Practically, a 1% increase in *central* corresponds to a small 0.166% decrease in crime rate. Notice that both Model 2 and 3 indicate that *prbarr*, *prbconv*, *polpc*, and *pctmin*80 are significant.

With respect to coefficient values, we see only minor shifts between Model 2 and 3. For a 1% increase in:

- *prbarr*, we see a shift from 0.373% (Model 2) to 0.344% (Model 3) in decrease in crime rate.
- *prbconv*, we see a shift from 0.310% (Model 2) to 0.294% (Model 3) in decrease in crime rate.

- *polpc*, we see a shift from 0.487% (Model 2) to 0.373% (Model 3) in increase in crime rate.
- *pctmin*80, we see a shift from 0.237% (Model 2) to 0.258% (Model 3) in increase in crime rate.

Comparatively, the $adjustedR^2$ also does not change much from Model 2 at 0.761 to Model 3 at 0.797. However, Model 3's AIC value, at 5.99, is much lower that of Model 2, at 19.5. This is likely due to the further removal of obs. 25, that is justified with explanation.

Chapter 5

Conclusion

5.1 Comparing Models

The three models are summarized in Table 5.1. Across all three models, *prbarr*, *prbconv*, and *polpc* (all three of which the mayor could reasonably influence) were found to be significant, with Model 2 and 3 adding *density* and *pctmin*80 as a highly significant variables. Notice here that the model fit as measured by AIC actually increase from model 1 to model 3. A portion of this decrease is caused by the removal of observations 77 and 25, as justified previously. From model 1 to model 2, most of the increased fit is not caused by the change in number of observation, but rather the additional explanatory variables *density* and *pctmin*80. On the other hand, the additional variables in model 3 were not highly significant, and thus the decrease in AIC is mostly caused by the removal of observation 25. Based on these findings, we suggest the model 2 actually offers the a balance between model fit and parsimony, despite the AIC values. The specific practical significance of each statistically significant variables were previously discussed in each model's respective sections.

Table 5.1: Linear Models of Crime Rates

	Dependent variable:		
	Base Model		
	(1)	(2)	(3)
log(prbarr)	-0.552^{***}	-0.373^{***}	-0.344^{***}
	(0.166)	(0.119)	(0.103)
log(prbconv)	-0.312^{*}	-0.310^{**}	-0.295^{***}
	(0.161)	(0.122)	(0.106)
log(prbpris)	0.228	−0.178	−0.067
	(0.359)	(0.200)	(0.159)
avgsen	−0.014	−0.018	-0.024^{*}
	(0.023)	(0.012)	(0.012)
log(polpc)	0.650^{***}	0.487^{**}	0.373^{*}
	(0.216)	(0.228)	(0.201)
log(density)		0.269^{***}	0.364^{***}
		(0.077)	(0.083)
log(wages)		0.041	0.362
		(0.511)	(0.521)
log(pctmin80)		0.237^{***}	0.259^{***}
		(0.039)	(0.066)
log(pctymle)		−0.006	−0.039
		(0.195)	(0.152)
taxpc			−0.005
			(0.006)
west			−0.014
			(0.147)
central			-0.166^{*}
			(0.091)
urban			−0.070
			(0.161)
Constant	0.055	−0.950	−3.132
	(1.608)	(3.722)	(3.527)
AIC	92.1	19.5	5.6
Observations	88	87	86
R^2	0.471	0.786	0.829
Adjusted R^2	0.438	0.761	0.798
Note:			*p<0.1; **p<0.05; ***p<0.01

5.2 Omitted Variable Bias

In this section we provide Table 5.2 listed some of the omitted variable in our models, the explanatory variables they confound, and the direction of the bias.

Omitted variable	Proxy	Bias
Average age of convicted criminal	avgsen, prbarr	Moderate bias towards from zero. Younger population are largely involved in petty crimes or misdemeanor can receive shorter sentences compared to felonies.
Average number of civilian owned firearms involved in crimes per county	avgsen, prbarr, prbconv	Moderate bias away from zero. More civilian owned firearms are involved in crimes, *avgsen* and the "probability" variables should all rise dramatically.
Exact location of crime	density, county	Strong bias away from zero. Many crimes might occur in the same neighborhood, and can be related events.
Neighborhoods	pctmin80, density, prbarr	Bias away from zero. Usually the same race/ethnicity lives in the neighborhood.
Location of police stations	prbarr	Small bias away from zero. High police density surrounding police stations, should lead to higher probability of arrest but lower crime rate.

Table 5.2: Omitted variables, respective proxies and biases.

5.3 Advised Policy

We propose, then, the campaign to focus on factors: *prbarr*, *prbconv*, *density*, and *pctmin*80. With arrests and convictions seemingly having a significant impact on crime rate, the campaign may want to consider policies that empower law enforcement organizations (with either resources, training, etc.) in these high population density areas. Specifically, our analysis suggests that focusing on increasing convictions would lead to an outsized effect on violent crime rates.

The campaign may also want to better understand the relationship between counties with higher minority populations and their crime rate. It may be beneficial, for example, to consider outreach programs and policies that create a level of trust and an open communication channel between law enforcement agents and minority communities.