Lab 3 - Reducing Crime

Clayton G. Leach, Karl I. Siil, Timothy S. Slade

July 23, 2018

Introduction

Our client is running for office in the state of North Carolina (NC). Her campaign commissioned us to research the determinants of crime in NC to help her develop her platform regarding crime-related policy initiatives at the level of local government. This report explores a 1994 dataset from Cornwell & Trumball that provides county-level economic, demographic, and crime data. Our analysis describes the dataset, presents initial summary statistics, and develops three linear regression models.

Initial Exploratory Data Analysis (EDA)

```
#here::dr_here() # TS: Use on first run to see why it fails on your machine, if it does.
here::here() # TS: Use when all set up, so we can mask the output.
#crime_raw <- read_csv('../crime_v2.csv', # TS: Commented out b/c using "here()"
crime_raw <- read_csv('crime_v2.csv', col_types = cols(prbconv = col_double()))
#problems(crime_raw) # TS: To comment out once we're ready to generate our draft report.
#crime_raw[97, ]

make_scatters <- function(df, var_list, y, trans) {

  if(!missing(trans)) {
    var_list <- append(var_list, str_glue('{trans}({var_list})'))
  }

  for (v in var_list){
  print(ggplot(df, aes_string(x = v, y = y)) +
          geom_point() +
          geom_smooth(method = 'lm', se = FALSE) +
          xlab(v) +
          ylab(y) +
          ggtitle(str_glue('{y} vs {v}'))
        )
    }
}
```

Missing Values

```
# KS: Rows with no data
crime_na <- crime_raw %>% filter_all(any_vars(!is.na(.)))
# KS: Row with one back tick
crime_na %>% filter_all(any_vars(is.na(.))) %>% select(which(!is.na(.)))
```

```
## # A tibble: 0 x 0
```

```
crime_na <- crime_na %>% filter_all(all_vars(!is.na(.)))
```

Upon loading the data, we examine the 6 rows that are missing data, finding that 5 are entirely blank and 1 contains only a backtick. We eliminate those to generate our working dataset.

Erroneous Duplicate Records

```
crime_na %>% count(county) %>% filter(n > 1) # county 193 is an exact duplicate
```

```
## # A tibble: 1 x 2
##   county     n
##    <int> <int>
## 1    193     2
```

```
#crime_na %>% filter(county == 193)
```

Continuing our QC, we note that one of the counties' records has been duplicated exactly. We therefore drop the duplicate record from our dataset.

```
crime_na <- crime_na %>% filter(!duplicated(.))
```

Plausibility Checks for Variables

Three of our key variables of interest (`prbarr`, `prbconv`, and `prbpris`) represent probabilities and should therefore theoretically be in the range of 0:1.

```
# look at weird 'probability' variables.
non_prob <- crime_na %>%
  filter(!between(prbarr, 0, 1) | !between(prbconv, 0, 1) | !between(prbpris, 0, 1))
```

Examining the data, we find 10 counties have values for the "probability" variables that are outside of the expected range. In each case, it is either `prbconv` (10 records) or `prbarr` (1 record) that fall outside the range.

Per the notes accompanying our data, *The probability of conviction is proxied by the ratio of convictions to arrests…* Given that definition, if not all suspects arrested are convicted, `prbconv` will be below 1. However, it may also exceed 1 if the number of exonerated suspects is exceeded by the number of suspects convicted of multiple charges. (See **here** for examples of multiple charges stemming from a single arrest.)

The notes on `prbarr` indicate *the probability of arrest is proxied by the ratio of arrests to offenses….* If multiple suspects are arrested for a single offense, and this happens more frequently than offenses which do not lead to arrests, `prbarr` would indeed exceed 1.

In both cases, there are plausible explanations for a probability value in excess of 1, however,one of the observations does appear to be an true outlier: The county labeled `115` has the lowest crime rate by far (~50% lower than that of any other county), the highest 'probability' of arrest (>1 arrest per offense, nearly 58% greater than the county with the second-highest probability), the longest average sentence (20.7 days, ~15% higher than the second-longest), and the largest number of police per capita (9 officers per 1,000 residents, more than twice as many as the second-highest county). While those numbers appear unusual, they are internally consistent: one would expect a very low crime rate from a county that has a very strong police presence, arrests a large proportion of suspects, and punishes convicted criminals severely. **However, given the extreme values across the board we will exclude this data point, as we do not believe it to be representative of the overall generative process.**

```
crime_na <- crime_na[crime_na$prbarr<=1,]
```

Examining the remainder of our data, we found no substantial evidence of *top-coded* or *bottom-coded* (i.e., truncated) variables which might bias our regression models. However, there is an extreme outlier in **wser**, the variable indicating the county's weekly wage in the service industry. To determine if this is valid we looked at the wage values for other sectors of the economy and did not see elevated values. It is improbable, but not impossible, that individuals in the service industry are making significantly more than anyone else in the county. We believe this data point should be scrutinized further before a determination is made to modify it.

Transformation Analysis

If the relationship between two variables is not linear, adding them to a linear regression model as-is (without a transformation) will generate inaccurate results and possibly result in an invalidation of our heteroskedsticity assumption. It is therefore important to explore whether the relationship between two variables shares some non-linear relationship and thus whether a transformation is required. As part of our EDA we explored this question for all of the variables in the dataset.

Our first step was to evaluate if any variables had signficant skew in their distributions by checking whether they generally conformed to a normal distribution using R's qqplot. While this is not necessarily a reason to transform a variable, it can help us identify variables of interest.

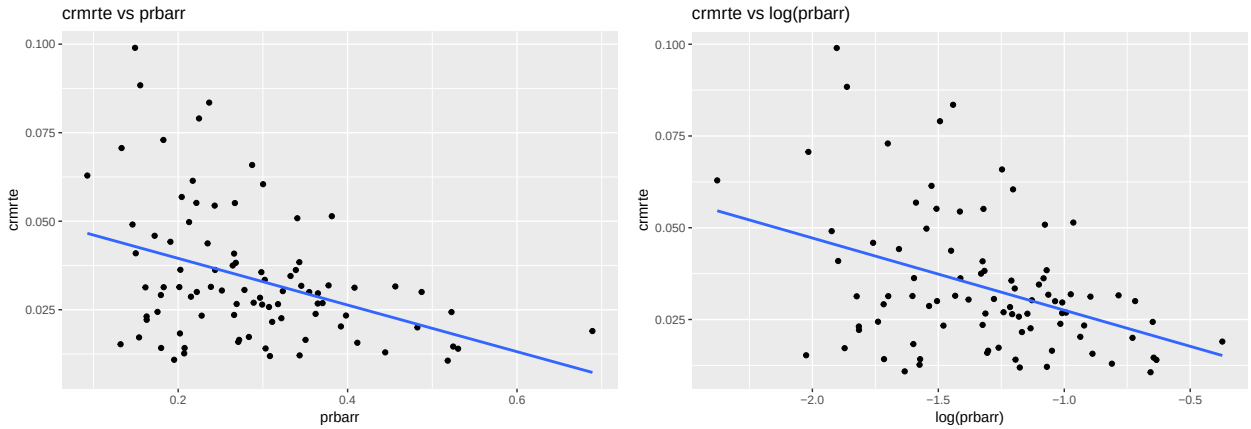
From our graphs we saw that probability of arrest (prbarr), probability of conviction (prbconv), police per capita (polpc), tax revenue per capita (taxpc), density (density),percent young male (pctymle), and mix (mix) deviated from normality. We will consider this in addition to other factors when deciding if a transformation would be beneficial.

Secondly, while not a perfect approach due to the possible interactions amongst variables, we also wanted to look at whether there is any obvious non-linearity when looking at crime rate and each variable independently. To do this we looked at a scatterplot of crime rate vs. each variable.

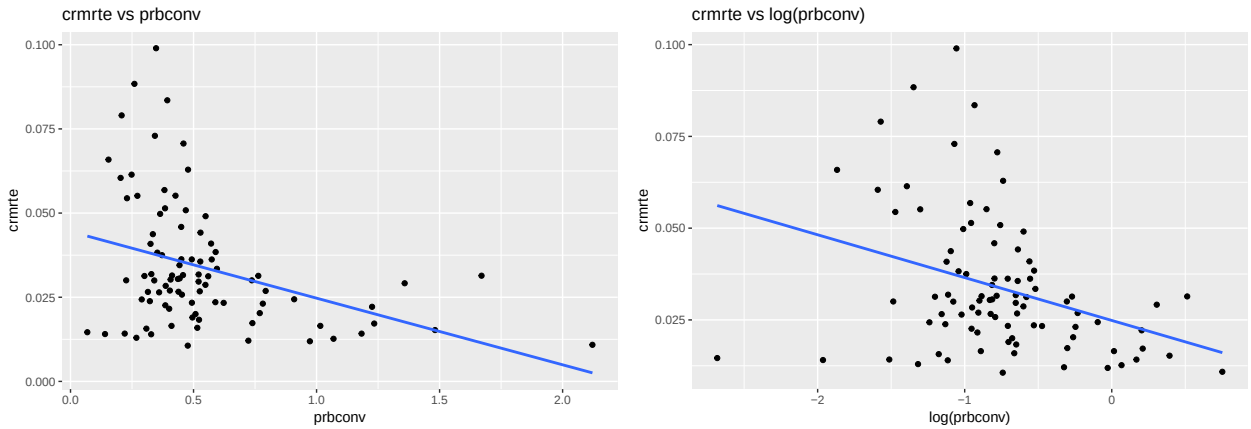
Reviewing these graphs yields five (5) variables which appear to have a non-linear relationship with crime rate: Probability of arrest(prbarr), probability of conviction (prbconv), police per capita (polpc), density (density), and tax revenue per capita (taxpc). Four of the variables appear to benefit from a log transform, while the fifth (density) appear to be related to crime rate via the square root function. In addition to checking whether the relationship appeared to improve in linearity, we also checked whether this transformation helped with variable skew. In every case the distribution of our variable moved closer to normality. Given the improvement in linearity and normality we will use these transformations moving forward.

Linearity Graphs:

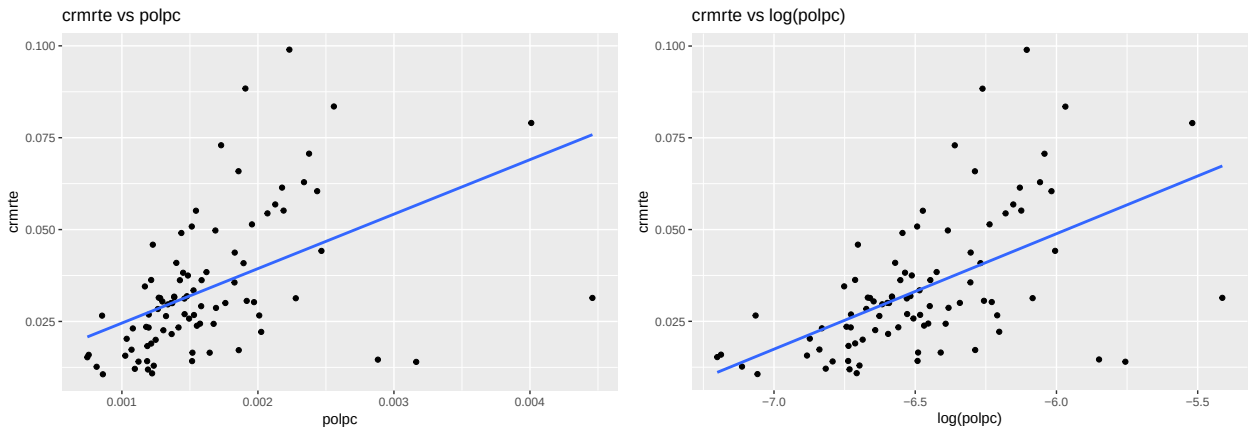
```
make_scatters(df = crime_na, var_list = c('prbarr'), y = 'crmrte', trans = 'log')
```



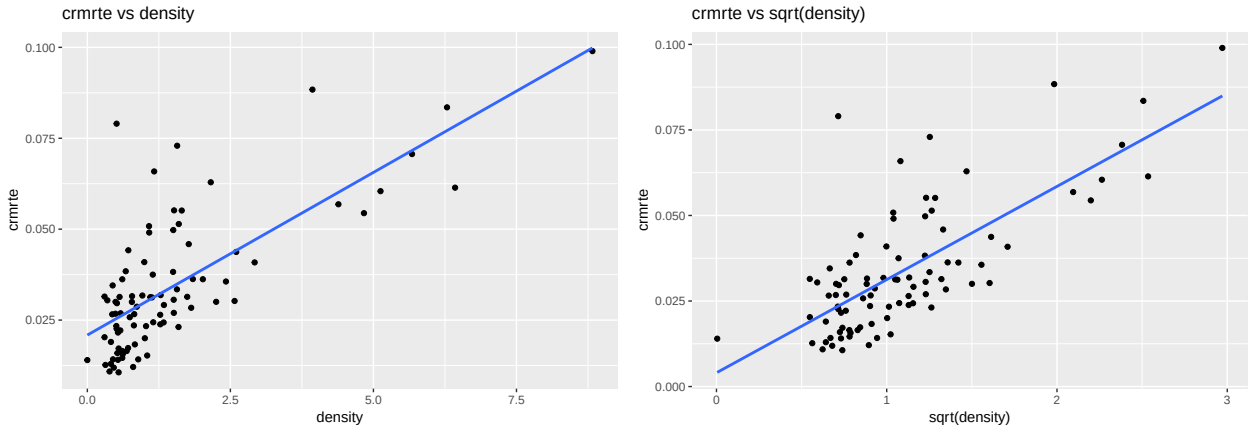

```
make_scatters(df = crime_na, var_list = c('prbconv'), y = 'crmrte', trans = 'log')
```



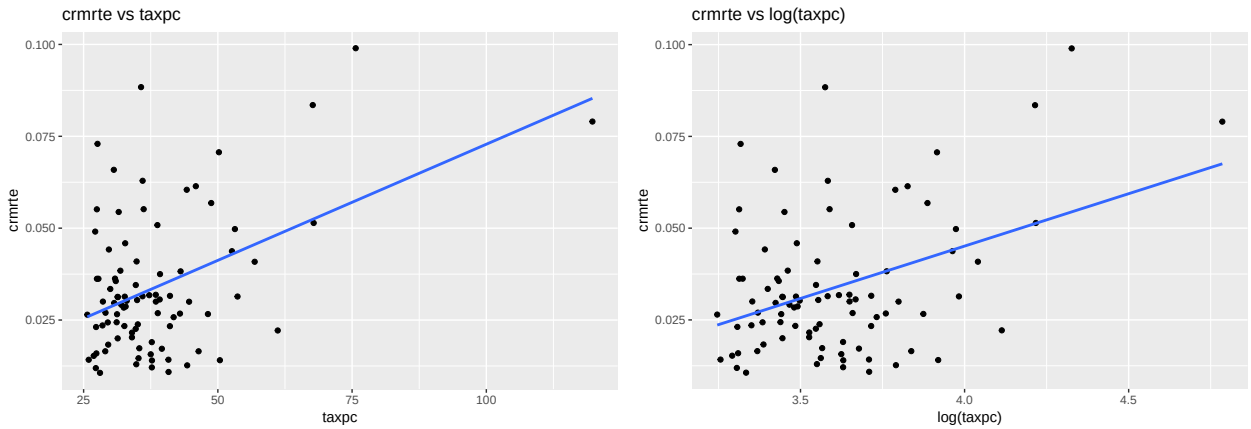

```
make_scatters(df = crime_na, var_list = c('polpc'), y = 'crmrte', trans = 'log')
```




```
make_scatters(df = crime_na, var_list = c('density'), y = 'crmrte', trans = 'sqrt')
```

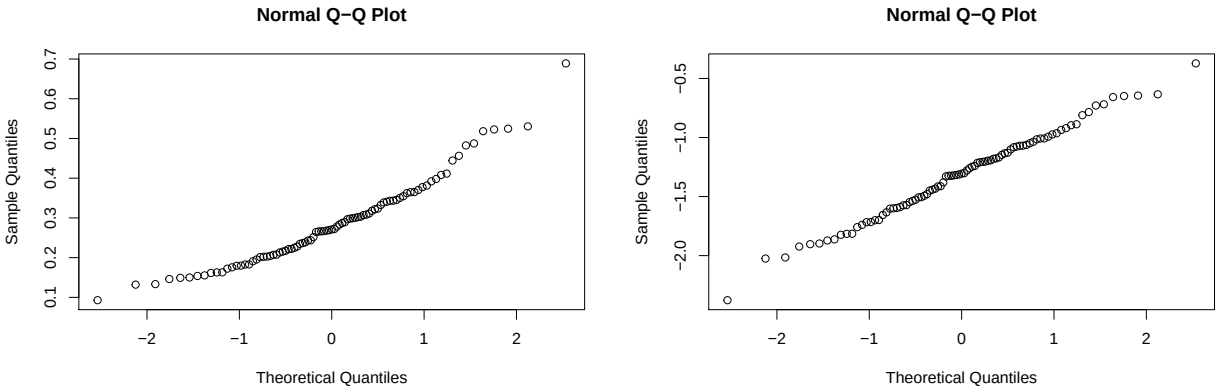



```
make_scatters(df = crime_na, var_list = c('taxpc'), y = 'crmrte', trans = 'log')
```

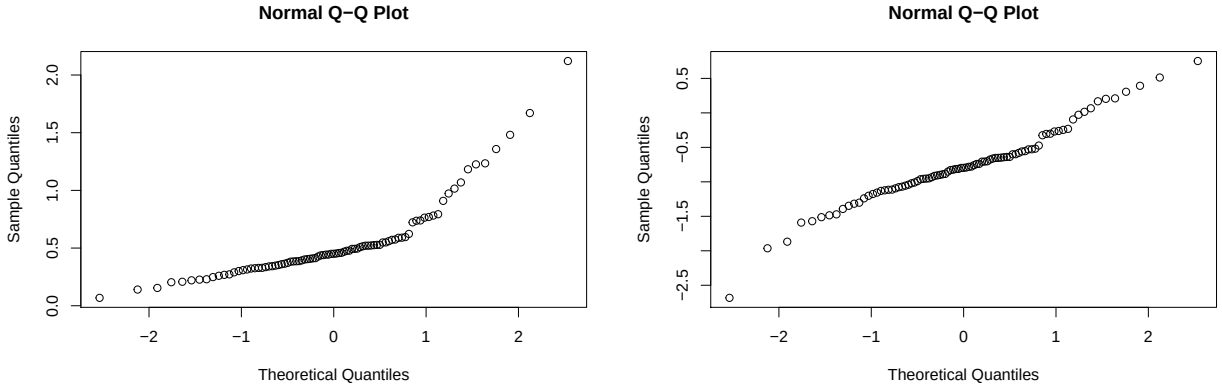


Normality/Skew With and Without Transformation:

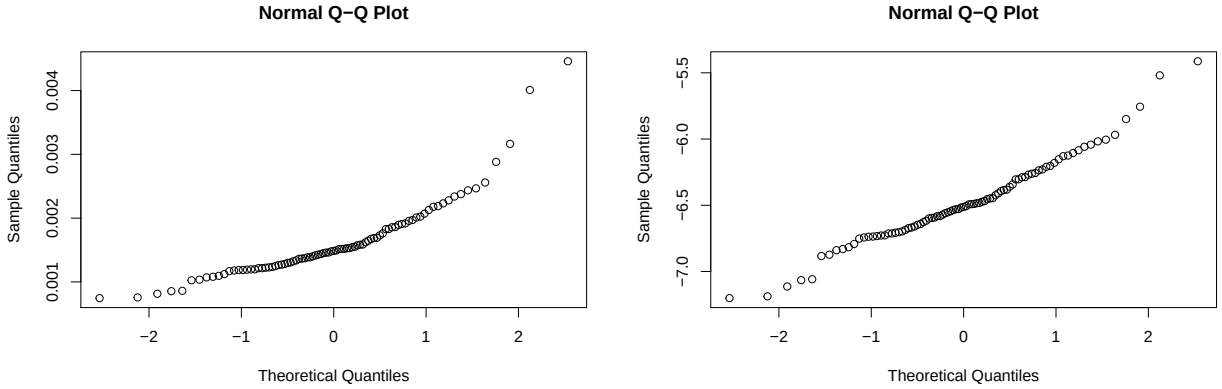
```
#Normality/skew with and without transformations
print(c(qqnorm(crime_na$prbarr), qqnorm(log(crime_na$prbarr))))
```



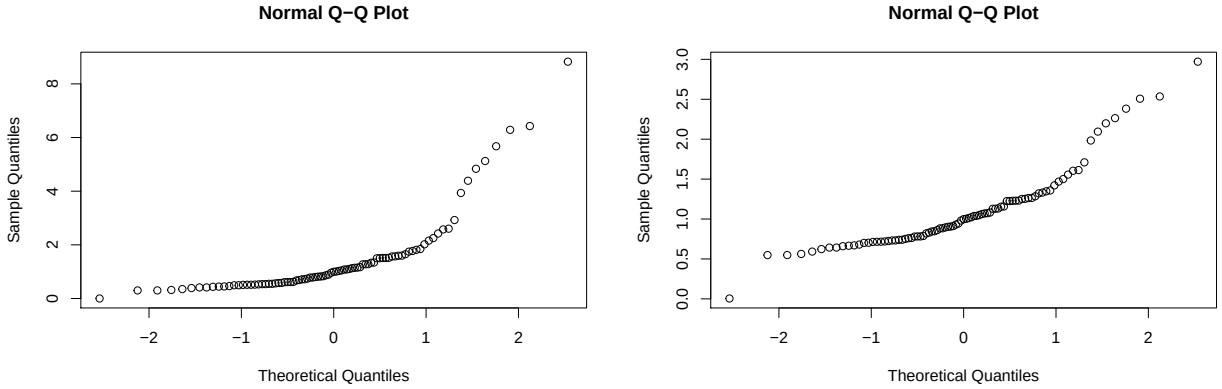

```
print(c(qqnorm(crime_na$prbconv), qqnorm(log(crime_na$prbconv))))
```



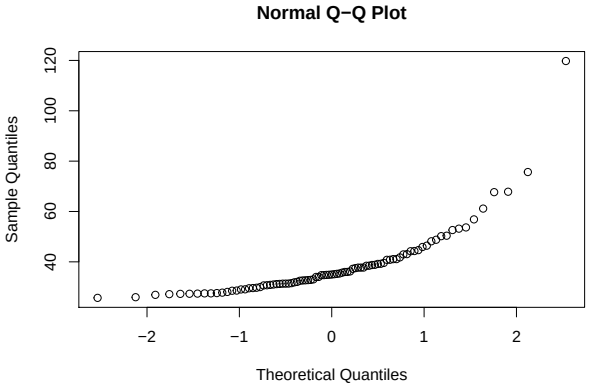

```
print(c(qqnorm(crime_na$polpc), qqnorm(log(crime_na$polpc))))
```

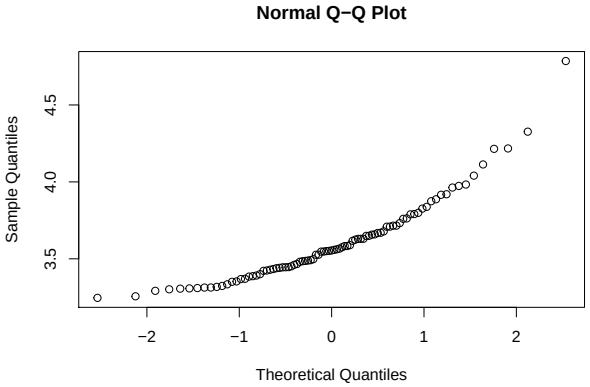



```
print(c(qqnorm(crime_na$density), qqnorm(sqrt(crime_na$density))))
```




```
print(c(qqnorm(crime_na$taxpc), qqnorm(log(crime_na$taxpc))))
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
40
60
80
100
120
−2
−1
0
1
2


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
3.5
4.0
4.5
−2
−1
0
1
2

Variables Available for Analysis

The table below details the variables available to us, which model(s) we included them in, and any transformations we applied before including them in our model.

Table 1: Hypothesized Primary Determinants of Observed Crime Rate

Variable Name	Description	Transformation Applied
`county`	*Source county of data*	-
`year`	*Source year of data*	-
`crmrte`	*crime rate*	-
`prbarr`	*'probability' of arrest*	-
`prbconv`	*'probability' of conviction*	-
`prbpris`	*'probability' of prison sentence*	square ($prbpris^2$)
`avgsen`	*average sentence, in days*	-
`polpc`	*police per capita*	$\log(polpc)$
`density`	*people per sq. mile*	$\sqrt{(density)}$
`taxpc`	*tax revenue per capita*	$\log(taxpc)$
`west`	*Dummy: source county of data is in Western NC*	-
`central`	*Dummy: source county of data is in Central NC*	-
`urban`	*Dummy: source county of data is urban*	-
`pctmin80`	*percent minority in 1980*	
`wcon`	*wages in the construction industry*	
`wtuc`	*wages in the transportation, utilities, and communication industries*	-
`wtrd`	*wages in the construction industry*	-
`wfir`	*wages in the finance, insurance, real estate industries*	-
`wser`	*wages in the service industry*	-
`wmfg`	*wages in the manufacturing industry*	-
`wfed`	*wages among federal employees*	-
`wsta`	*wages among state employees*	-
`wloc`	*wages among local government employees*	-
`mix`	*mix of offenses; face-to-face v others*	-
`pctymle`	*percent young male*	$\log(pctymle)$

Research Question and Model-Building

Our **research question** is the following: **Should our candidate support a traditional "Tough on Crime" platform?**

We face a key limitation: our data does not give us visibility into the crimes themselves or changes in crime, but rather provides only the official *crime rate*. The crime rate is a function not only of crimes committed but also of various factors, some of which may be unobservable. For instance, poor community-police relations may bias crime rates downward if an area's residents **do not report all the crimes they observe or experience**. Conversely, those poor relations may also bias crime rates upward if police officers engage in **predatory policing practices** and the community lacks the wherewithal to fight back. A full discussion of omitted variable bias will occur later in this analysis.

In order to answer our research question we created several models which included variables related to key crime policy decisions. We only included variables which would allow our canadidate to make concrete policy proposals that lie within her purview.

Using the aforementioned criteria we choose the variables police per capita (`polpc`), probability of arrest (`prbarr`), probability of conviction (`prbconv`), probability of incarceration (`prbpris`), and average sentence length (`avgsen`) for our first model. Understanding how these items relate to crime rate can help shape her position, and understand whether a "Tough on Crime" stance does in fact acheive a reduction in crime. Specifically, we can help her understand which departement/function should receive additional funding given a limited budget (e.g. if conviction rates are highly correlated perhaps we invest more in our District Attorneys).

Table 2: Model 1: Hypothesized Key Determinants of Observed Crime Rate

Variable Name	Description	Transformation Applied
`prbarr`	*'probability' of arrest*	-
`prbconv`	*'probability' of conviction*	-
`prbpris`	*'probability' of prison sentence*	square ($prbpris^2$)
`avgsen`	*average sentence, in days*	-
`polpc`	*police per capita*	log(*polpc*)

```
#Linear Regression model using only our key variables of interest, transformed as needed
model1_trans <- with(crime_na, lm(crmrte ~ log(polpc)+log(prbarr)+log(prbconv)+prbpris+avgsen))

#Adding AIC to our model to help us compare models in the future.
model1_trans$AIC <- AIC(model1_trans)

#Output model results in nice format using tidy and kable
kable(tidy(model1_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.1402673	0.0373427	3.7562149	0.0003193
log(polpc)	0.0224426	0.0049818	4.5049668	0.0000215
log(prbarr)	-0.0226831	0.0041956	-5.4064160	0.0000006
log(prbconv)	-0.0126558	0.0029704	-4.2606258	0.0000535
prbpris	0.0115442	0.0180965	0.6379269	0.5252773
avgsen	-0.0005701	0.0005866	-0.9717486	0.3339991

Comments on Model 1:

The log of police per capita, the log of the probability of arrest, and the log of probability of conviction all have high levels of significance. The overall model has an adjusted R^2 of 0.486. There are three main points to highlight:

1) Our coefficient for log police per capita (`polpc`) is positive and highly significant. If we inaccurately assumed this model was causal the best mechanism to reduce crime rates would be to sunset the police force! However, a more plausible interpretation is that a higher number of police per capita is a response to higher levels of criminal activity, rather than a cause of it.

2) Both log probability of arrest (`prbarr`) and log probability of conviction (`prbconv`) have highly significant and negative coefficients. This fits with what we would expect: the more likely an individual is to be caught and convicted the less likely they are to commit crime.

3) Neither the probability of incarceration of the average sentence length have high levels of significance. Furthermore, the probability of prison has a positive coefficient, indicating that criminal activity increases are correlated with higher incarceration rates (ceteris perebus). This is highly counterintuitive if we were to interpret this causally.

Model 2

While our first model showed promise there are several other factors which might be correlated with these explanatory variables; this would lead to multicollinearity issues. When multiple independent variables exhibit collinearity it becomes increasingly difficult to untangle their individual effects on the crime rate. This is troublesome because the model coefficients may misrepresent the impact of a variable, leading to a policy which has underwhelming impact. In order to avoid biased coefficient estimates, and to improve upon our model we created second model which includes variables we believe to be highly correlated with our three key variables.

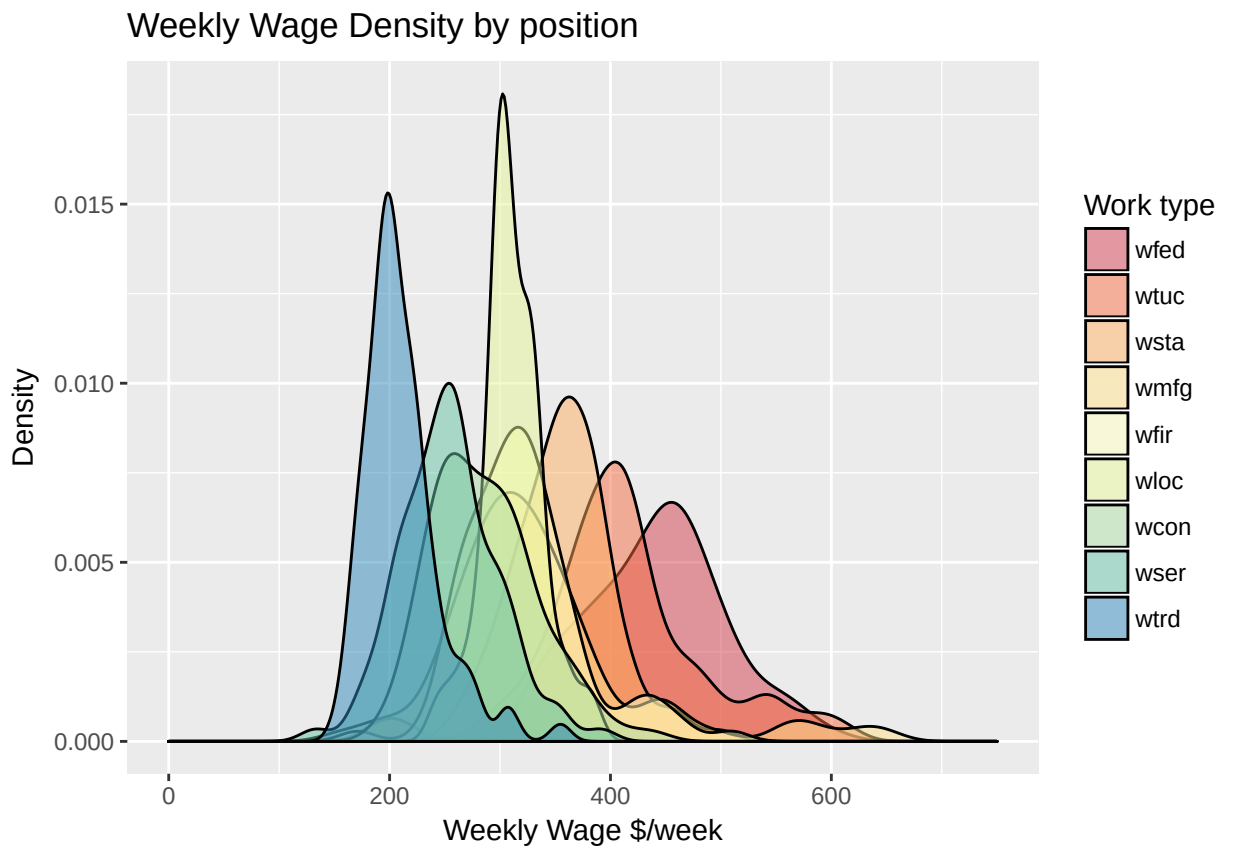
Collinearity with Police Per Capita: We would expect policing practices in urban areas to differ substantially from those in suburban or rural areas; including density can help control for this. Additionally, the police force is funded by taxpayers, and therefore we might expect a larger police force per capita to be correlated with higher tax revenues.

Creation of New Variables To Simplify Wage Metrics

Our dataset contains 9 wage variables, each representing a different sector or group of industries. We do not have *a priori* justification to believe *a single industry* might contribute disproportionately to crime, but we can assume that low wages in general might create an environment of economic scarcity in which crime incidence would increases. Including all 9 variables when our dataset only contains 90 observations would be extremely limiting, but excluding them entirely prohibits us from understanding how microeconomic conditions may be contributing to observed crime rates. Researching the composition of each county's economy and weighting each variable accordingly might be a fruitful strategy, but it lies outside the scope of this report. The solution we ultimately implemented was to create three new composite variables:

1) Government Wage: Average of `wfed` (federal government wage), `wsta` (state government wage), and `wloc` (local government wage)
2) Blue-Collar Wage: Average of `wmfg` (manufacturing), `wser` (service), `wcon` (construction)
3) Professional Wage: Average of `wfir` (Finance/Investment/Real Estate), `wtrd` (Wholesale/Retail Trade), and `wtuc` (Transportation, Utilities, Communication)

```
crime_na %>%
  gather(wfed, wtuc, wsta, wmfg, wfir, wloc, wcon, wser, wtrd, key = 'work_type',
         value = 'weekly_wage') %>%
  select(county, work_type, weekly_wage, everything()) %>%
  mutate(work_type = factor(work_type, c('wfed', 'wtuc', 'wsta', 'wmfg', 'wfir',
                                         'wloc', 'wcon', 'wser', 'wtrd'))) %>%
  ggplot(aes(x = weekly_wage, fill = work_type)) +
  geom_density(alpha = 0.5) +
  scale_fill_brewer(palette = 'Spectral') +
  ggtitle('Weekly Wage Density by position') +
  xlab('Weekly Wage $/week') +
  ylab('Density') +
  labs(fill = 'Work type') +
  xlim(0, 750)
```




```
crime_na$govt_wg <- (crime_na$wfed+crime_na$wsta+crime_na$wloc)/3
crime_na$physical_wg <- (crime_na$wmfg+crime_na$wser+crime_na$wcon)/3
crime_na$industry_wg <- (crime_na$wfir+crime_na$wtrd+crime_na$wtuc)/3
```

Collinearity with Probability of Arrest: Sociological research suggests that men - especially young, minority men - are at an increased risk of arrest. Therefore we will include both the `pctymle` (percentage of young male, under a log transformation) and `pctmin80` (percentage of minorities in 1980) variables.

Collinearity with Probability of Conviction: The likelihood of an arrest leading to a conviction depends not only on the culpability of the suspect, but also on the quality or effectiveness of the police department's investigative team, the district attorney, court-appointed advocates, judges, and other government officials. Quality is an unobservable (omitted) variable, but it might be loosely correlated with overall government wages. To proxy this we will include our government wage (govt_wg) variable.

Collinearity with Probability of Incarceration and Average Sentencing: While there may be unobservable covariates for these two variables, we cannot identify any reasonable proxies to include in a revised model to address the issue.

Table 4: Model 2: Hypothesized Key Determinants of Observed Crime Rate with Additional Covariates

Variable Name	Description	Transformation Applied
`prbarr`	*'probability' of arrest*	-
`prbconv`	*'probability' of conviction*	-
`prbpris`	*'probability' of prison sentence*	square ($prbpris^2$)
`polpc`	*police per capita*	$\log(polpc)$
`avgsen`	*average sentence, in days*	-
`taxpc`	*tax revenue per capita*	$\log(taxpc)$
`density`	*persons per square mile*	$\sqrt{density}$
`pctymle`	*percentage of young males*	$\log(pctymle)$

```r
#Model with our key explanatory variables, and what we suspect to be key covariates
model2_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) +
                                  log(prbconv) + prbpris + avgsen + log(taxpc) + sqrt(density) +
                                  govt_wg + pctymle + pctmin80))

model2_trans_no_minvar <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) +
                                  log(prbconv) + prbpris + avgsen + log(taxpc) + sqrt(density) +
                                  govt_wg + pctymle))

added_adj_r_squared <- summary(model2_trans)$adj.r.squared - summary(model2_trans_no_minvar)$adj.r.squa

#Adding AIC to our model to help us compare models in the future.
model2_trans$AIC <- AIC(model2_trans)

#Output model results in nice format using tidy and kable
kable(tidy(model2_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0313065	0.0445102	0.7033550	0.4839288
log(polpc)	0.0111145	0.0039994	2.7790322	0.0068286
log(prbarr)	-0.0159101	0.0033446	-4.7568992	0.0000089
log(prbconv)	-0.0102757	0.0021993	-4.6723480	0.0000122
prbpris	-0.0100601	0.0124888	-0.8055300	0.4229628
avgsen	-0.0004538	0.0003962	-1.1452575	0.2556043
log(taxpc)	0.0066255	0.0045139	1.4678087	0.1461772
sqrt(density)	0.0189635	0.0028549	6.6424115	0.0000000
govt_wg	-0.0000113	0.0000419	-0.2706261	0.7873934
pctymle	0.0448356	0.0467876	0.9582797	0.3408827
pctmin80	0.0003801	0.0000620	6.1316223	0.0000000

Comments on Model 2:

1) One item of interest was the extreme degree of significance we see associated with our percent minority variable (`pctmin80`). Interestingly, when using only this variable to predict crime rates our R^2 is very low; after controlling for other factors, however, this variable becomes extremely important. One way to measure this importance is by calculating the difference between the adjusted R^2 values for a model that includes the variable and one that excludes it. When including the minority variable in our model, the adjusted R^2 is 0.769. When excluding it, the adjusted R^2 is 0.662, a difference of 0.107

2) The square root of the population density is also highly significant. Given the transformation the correct interpretation is that crime activity increases quickly when moving from very small to medium sized population densities, but requires increasingly large levels of population increase to have an effect on crime rate.

Model 3

Given the countless ways behavioral issues are interconnected, we wondered whether every variable we had data on might be correlated with either the crime rate or an already included variable in some fashion. Our focus was to determine if including all our variables substantially changed the significance or coefficient of any of our previously included variables. Additionally, we wanted to understand if any of the variables we had previously left out were in fact predictive overall.

```r
#Linear model including our key explanatory variables, suspected covariates, and most other variables
model3_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) + prbpris +
                            avgsen + log(taxpc) + sqrt(density) + govt_wg + pctymle +
                            pctmin80 + west + central + urban +
                            physical_wg + industry_wg + mix))

#Adding AIC to our model to help us compare models in the future.
model3_trans$AIC <- AIC(model3_trans)

#Output model results in nice format using tidy and kable
kable(tidy(model3_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0738841	0.0529469	1.3954385	0.1671727
log(polpc)	0.0144337	0.0043432	3.3232644	0.0014010
log(prbarr)	-0.0142369	0.0035268	-4.0367943	0.0001336
log(prbconv)	-0.0093405	0.0023181	-4.0293967	0.0001371
prbpris	-0.0057449	0.0126456	-0.4543019	0.6509782
avgsen	-0.0009229	0.0004212	-2.1908633	0.0316994
log(taxpc)	0.0045176	0.0052726	0.8568070	0.3943944
sqrt(density)	0.0177847	0.0041576	4.2776578	0.0000572
govt_wg	-0.0000142	0.0000448	-0.3159036	0.7529896
pctymle	0.0324232	0.0473981	0.6840610	0.4961325
pctmin80	0.0003037	0.0000984	3.0849009	0.0028880
west	-0.0055698	0.0040533	-1.3741343	0.1736626
central	-0.0047271	0.0028764	-1.6433712	0.1046664
urban	0.0029053	0.0059482	0.4884266	0.6267320
physical_wg	-0.0000260	0.0000158	-1.6464321	0.1040333
industry_wg	0.0000330	0.0000319	1.0360485	0.3036467
mix	-0.0169111	0.0160818	-1.0515676	0.2965145

Model 3 Notes:

Similarly to model 2, we find that log police per capita, log probability of arrest, log probability of conviction, square root of density, and percent minority are significant. While model 3 coefficients do change relative to model 2, there are not any drastic changes, or sign (+/-) changes indicating a reversal of effect.

Our adjusted R squared value for this model is 0.78 which does represent a slight, albeit negligible improvement over our previous model.

Model 4

One item we wished to investigate is whether we could build a more parsimonious model by selectively removing variables from our second model which did not appear to be significant. Note: this is a form of model "dredging" and these results are for comparative purposes only.

```r
#Linear model including our key explanatory variables, suspected covariates, and most other variables
model4_trans <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
                                  sqrt(density) + pctmin80 ))

#Adding AIC to our model to help us compare models in the future.
model4_trans$AIC <- AIC(model4_trans)
```

```
#Output model results in nice format using tidy and kable
kable(tidy(model4_trans))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.0572596	0.0239222	2.393576	0.0189401
log(polpc)	0.0129764	0.0033453	3.878948	0.0002093
log(prbarr)	-0.0165693	0.0031447	-5.268999	0.0000011
log(prbconv)	-0.0110556	0.0020958	-5.275054	0.0000010
sqrt(density)	0.0184496	0.0023137	7.974101	0.0000000
pctmin80	0.0003875	0.0000590	6.568238	0.0000000

Model Comparison

Below is a comparison table for our three initial models, as well as the fourth model that excluded all non-significant predictors. The table reports key statistics related to each model, including the coefficients for each predictor, the R^2 and adjusted R^2, and the *Akaike information criterion*, or AIC.

Our findings match with what we would hope to see from a model building perspective: Our AIC and adjusted R^2 numbers suggest that of our three original models, *Model 2* (which includes both our key explanatory variables and plausible covariates) performs the best. Our model which excludes key covariates has significantly less predictive power (as measured by adjusted R^2), and our model which includes everything–despite having the highest R^2 values–performs slightly less well on the AIC (a measure of explanatory power and parsimony).

```
stargazer(model1_trans, model2_trans, model3_trans, model4_trans,
          type = "latex", report="vc", header=FALSE,
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("aic", "rsq","adj.rsq","n"), omit.table.layout = "n")
```

Omitted Variables

Despite the promising results from our three models it is difficult to ascribe causality to the variables of interest. One issue with causal inference in general is ommitted variable bias, which can invalidate our ability to assume each explanatory variable is uncorrelated with the error term. While there are infinite variables which exist, there are several which deserve commentary:

1) Political Party in Control: Traditionally the issue of crime policy is a highly partisan issue, with each party having very different approaches to crime reduction. All else being equal, we might expect police levels and average sentence lengths to be correlated with the party that is crafting legislation. Assuming we construct our variable as a boolean indicator ("is_republican"), we might expect the coefficient for police per capita to diminish as we assume a priori that higher police levels are correlated with conservative crime policies. We could include this data by appending public records to our dataset which would be more appropriate than trying to find some other variable to proxy.

2) Unemployment Rate: While we have data on weekly wages, this does nothing to tell us what percentage of the population was actually earning those wages. It is likely that a higher unemployment rate would be correlated with higher rates of crime as people who may not normally commit criminal activity are pushed to their limits. We might also wish to be more granular, and include both minority and majority unemployment rates to help control for racial inequality. We realistically could obtain this data from the Bearue of Labor Statistics and append it to our dataset; we leave this next step to future researchers.

Table 8: Linear Models Predicting Crime Rate

	Dependent variable:			
	crmrte			
	(1)	(2)	(3)	(4)
log(polpc)	0.022	0.011	0.014	0.013
log(prbarr)	−0.023	−0.016	−0.014	−0.017
log(prbconv)	−0.013	−0.010	−0.009	−0.011
prbpris	0.012	−0.010	−0.006	
avgsen	−0.001	−0.0005	−0.001	
log(taxpc)		0.007	0.005	
sqrt(density)		0.019	0.018	0.018
govt_wg		−0.00001	−0.00001	
pctymle		0.045	0.032	
pctmin80		0.0004	0.0003	0.0004
west			−0.006	
central			−0.005	
urban			0.003	
physical_wg			−0.00003	
industry_wg			0.00003	
mix			−0.017	
Constant	0.140	0.031	0.074	0.057
Observations	89	89	89	89
R^2	0.515	0.795	0.820	0.782
Adjusted R^2	0.486	0.769	0.780	0.769
Akaike Inf. Crit.	−506.645	−573.434	−572.626	−577.871

3) Concentrated/Siloed Urban Blight: Our data is at the county level and therefore may obscure differences within the county. We would expect there to be a difference in crime rates between a county which is relatively homogenous with respect to the variables, and one which has drastic differences (e.g. a very poor area and a very nice area). One way to proxy this might be to calculate a normalized standard deviation of housing prices which could help capture if this phenomenom exists.

4) Policing Methodology: In recent years there has been a growing focus on "warrior" versus "guardian" mindsets in policing. Depending on the type of methodology we might see very different rates of arrest and conviction.

5) Police Representation: A community's relationship with the police can vary drastically from place to place. In recent years certain cities have had significant unrest, with a key element being a majority white police force existing in a largely minority community. Similar to police methodology, we might expect very different rates of arrest and conviction depending on whether the community feels the police represent them and their interests.

Conclusion: Findings and Policy Recommendations

Unfortunately, we only have a single cross section of data at our disposal and therefore it is nearly impossible to determine whether our variables are causes or effects. In reality it is most likely a combination of both, with changes in crime rate driving changes in policy, which in turn impact crime rates, ad infinitum. Because our ability to conduct causal inference is restricted, it would be unwise to make specific policy reccomendations based on our model.

We can recommend to our candidate that she advocate for increased investment in research that will investigate the relationships between crime rates and the following factors:

1) The number of police per capita
2) The likelihood that crimes, once reported, result in arrests
3) The likelihood that arrests result in convictions
4) The population density in a given area
5) The demographic composition in a given area

The relationships suggested by our models are intuitive: an increase in the number of police per capita and the population density in an area are associated with (predictive of, in our model) increases in the crime rate; increases in the probability of arrest and probability of conviction are associated with decreases in the crime rate. An increase in the percentage of minority residents is associated with an increase in the crime rate.

Four of them are simple to interpret: a 1% increase in the number of police is associated with an increase in the crime rate of 11.1 crimes per 100,000 people. Increasing by 1% the probability that a crime report leads to an arrest is associated with a decrease in the crime rate of 15.9 crimes per 100,000 people; a similar increase in the probability of conviction is associated with a decrease in the crime rate of 10.2 crimes per 100,000 people. A 1% increase in the minority share of the population is associated with an increase of 3.8 crimes per 100,000 people. The relationship with the population density is more complex, however, and bears further investigation