Lab 3 - Reducing Crime

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

August 6, 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 subset of the county-level data from Cornwell & Trumball's *Estimating the Economic Model of Crime with Panel Data (1994)* that provides various economic, demographic, and crime indicators for 1987. Our analysis describes the dataset, presents initial summary statistics, develops several linear regression models, and proposes additional research to inform policy recommendations.

Initial Exploratory Data Analysis (EDA)

We begin by exploring our dataset. We see that it has 97 records and 25 variables.

The notes we receive provide the following insight into the variables:

Table 1: Available Variables from Cornwell & Trumball (1994)

#	Variable Name	Type	Description
1	`county`	integer	*Source county of data*
2	`year`	integer	*Source year of data*
3	`crmrte`	numeric	*crime rate*
4	`prbarr`	numeric	*'probability' of arrest*
5	`prbconv`	numeric	*'probability' of conviction*
6	`prbpris`	numeric	*'probability' of prison sentence*
7	`avgsen`	numeric	*average sentence, in days*
8	`polpc`	numeric	*police per capita*
9	`density`	numeric	*people per sq. mile*
10	`taxpc`	numeric	*tax revenue per capita*
11	`west`	dummy	*source county of data is in Western NC*
12	`central`	dummy	*source county of data is in Central NC*
13	`urban`	dummy	*source county of data is urban*
14	`pctmin80`	numeric	*percent minority in 1980*
15	`wcon`	numeric	*wages in the construction industry*
16	`wtuc`	numeric	*wages in the transportation, utilities, and communication industries*
17	`wtrd`	numeric	*wages in the construction industry*
18	`wfir`	numeric	*wages in the finance, insurance, real estate industries*
19	`wser`	numeric	*wages in the service industry*
20	`wmfg`	numeric	*wages in the manufacturing industry*
21	`wfed`	numeric	*wages among federal employees*
22	`wsta`	numeric	*wages among state employees*
23	`wloc`	numeric	*wages among local government employees*
24	`mix`	numeric	*mix of offenses; face-to-face v others*
25	`pctymle`	numeric	*percent young male*

We see some variables which *a priori* seem useful: the 'probability' variables, police per capita, tax revenue, wages, and youth and minority composition of a county. Before exploring them further, however, we search the entire dataset for missing values that may affect our analyses.

Missing Values

We find 6 rows that are missing data; further scrutiny shows 5 are entirely blank and 1 contains only a backtick. We eliminate those to generate our working dataset.

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

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

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

Erroneously Duplicated Records

Continuing our QC, we note that 1 of the counties' records has been duplicated exactly. We therefore drop the duplicate record from our dataset.

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

county	n
193	2

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

Plausibility Checks for Variables

We see from the notes that three of our key variables of interest (`prbarr`, `prbconv`, and `prbpris`) represent probabilities and should therefore theoretically be in the range of 0:1.

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

Examining the data,[1] 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 appears to be an outlier. The county labeled `115` has the lowest crime rate by far (~50% lower

[1] See *Appendix* for the `non_prob` table

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 also 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.[2]

Check for Truncated Variables and Outliers

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.

We see a handful of outliers. In `wser`, the variable indicating the county's weekly wage in the service industry, we find that county 185 is an extreme outlier. To determine if this is valid we looked at the wage values in that county for other sectors of the economy.

```
# Analysis of outliers
outlier_wser <- filter(crime_na, crime_na$wser == max(crime_na$wser))
outlier_wser[, c(1, 15:ncol(outlier_wser))]
```

```
## # A tibble: 1 x 12
##   county  wcon  wtuc  wtrd  wfir  wser  wmfg  wfed  wsta  wloc    mix
##    <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1    185  227.  332.  167.  264. 2177.  248.  381.  367.  300. 0.0497
## # ... with 1 more variable: pctymle <dbl>
```

```
mean_other_wage <- with(outlier_wser, mean(wcon, wtuc, wtrd, wmfg, wfed, wsta,
    wloc))
```

It is improbable, but not impossible, that individuals in the service industry are making 9.6X more than the average wage of all other workers in the county. It seems more likely that a keystroke error was made in the recording of this variable (e.g., 2177.07 → 217.71). As we get into model-building we will check whether this particular value exerts undue influence on our results.

The other "outliers" we see are the values associated with county `115` discussed above, but those are neither as extreme nor inconsistent with theory.

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 heteroskedasticity and zero conditional mean assumptions. 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 significant 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 helped us identify variables of interest. Below we present the code we used to generate the series of graphs as well as two sample graphs for illustrative purposes.[3]

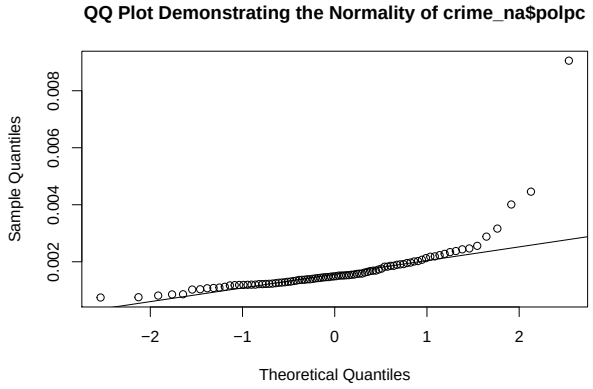
```
make_qqs <- function(df, var_list, trans) {
    var_list_new <- c()
```

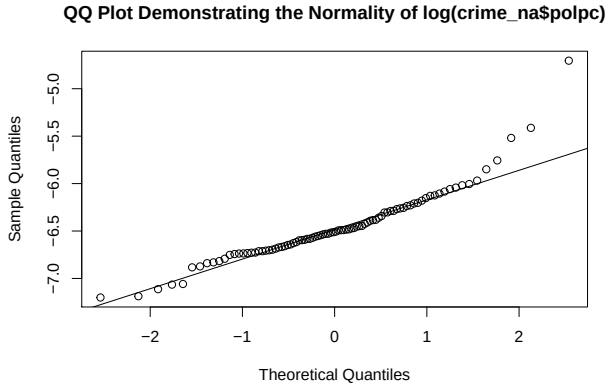
[2]Discussion with peers and further research reveals that the county labels are, in fact, FIPS (Federal Information Processing Standard) codes. A deep dive into additional county-level contextual factors that could inform our analyses is beyond the scope of this report; we leave it as an exercise for the reader.

[3]See *Appendix* for the remainder of the graphs.

```
    for (var in var_list) {
        var_list_new <- append(var_list_new, str_glue("{df}${var}"))

        if (!missing(trans)) {
            var_list_new <- append(var_list_new, str_glue("{trans}({df}${var})"))
        }
    }

    for (col in var_list_new) {
        title <- str_glue("QQ Plot Demonstrating the Normality of {col}")
        qqnorm(eval(parse(text = col)), main = title)
        qqline(eval(parse(text = col)))
    }
}
```

```
make_qqs("crime_na", "polpc", "log")
```





From our graphs we saw that probability of conviction (`prbconv`), police per capita (`polpc`), probability of arrest (`prbarr`), tax revenue per capita (`taxpc`), population density (`density`), proportion of young males in the population (`pctymle`), and the mix of face-to-face vs. impersonal crimes (`mix`) all deviated from normality. We will consider this in addition to other factors when deciding if a transformation would be beneficial.

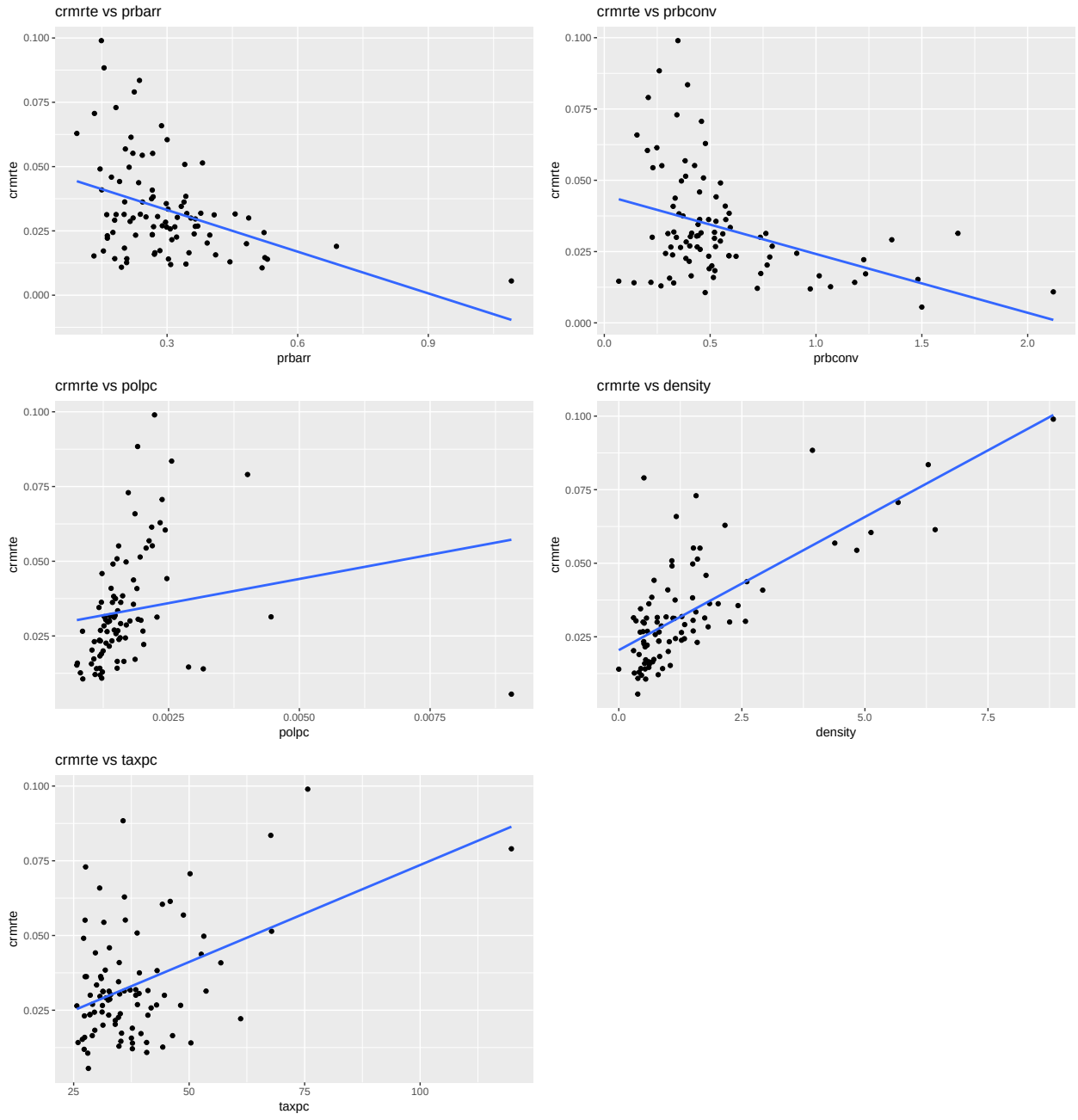
Checking for Linearity Between Predictor and Response Variables

While it is not a perfect approach due to the possible interactions amongst independent 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, and then applied various transformations to see whether those would improve the distribution. We present sample code and an illustrative graph here as an example.[4]

```
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(substitute(y)) + ggtitle(str_glue("{y} vs {v}")))
```

[4]See *Appendix* for remaining graphs.

```
    }
}
make_scatters(crime_raw, c("prbarr", "prbconv", "polpc", "density", "taxpc"),
    "crmrte")
```

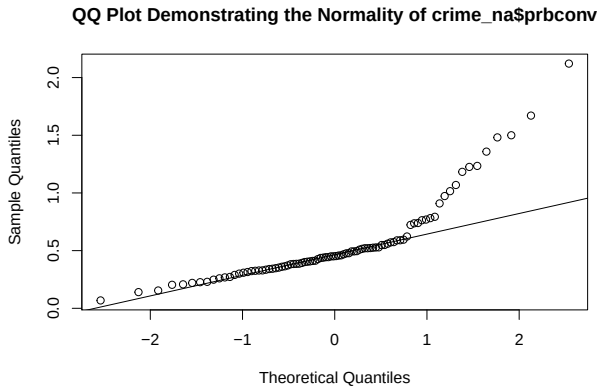


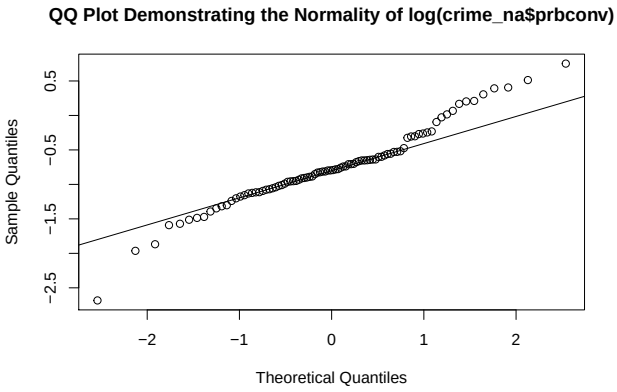
Reviewing these graphs yields five (5) variables which appear to have a non-linear relationship with crime rate: `prbarr`, `prbconv`, `polpc`, `density`, and `taxpc`. While it is not particularly important that predictor variables be normally distributed, we do want them to display both *symmetry* and *high variance*; to the extent applying a transformation advances those goals, it is worth considering provided it does not inhibit interpretability or run strongly counter to theory.

Four of the variables appear to benefit from a log transform, while the fifth (`density`) appears to be related to crime rate via the square root function. The pre- and post-transformation scatterplots and q-q plots for

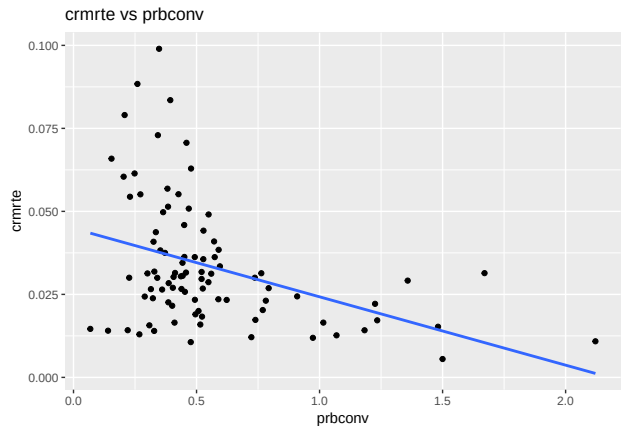
prbconv are presented here for illustrative purposes.[5]

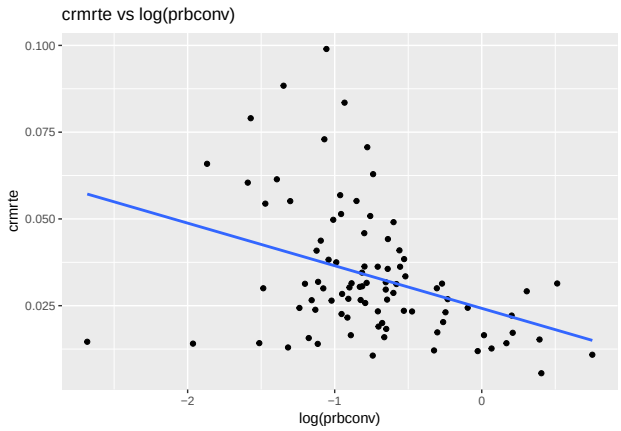
```
make_qqs(df = "crime_na", var_list = "prbconv", trans = "log")
```






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



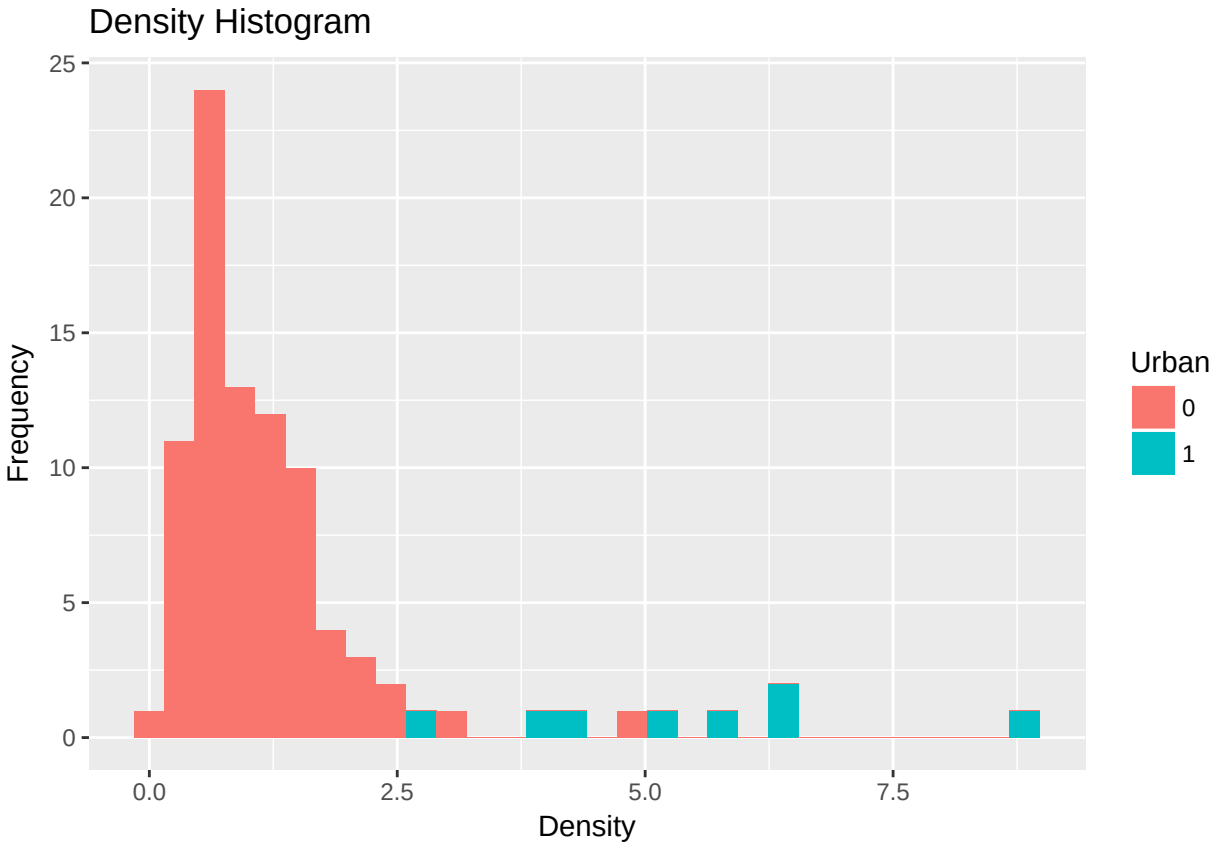


In addition to checking whether the relationship appeared to improve in linearity, we also checked whether this transformation helped with variable skew. In every case the distribution of our variable moved closer to normality.

With regards to interpretability, the interpretation of a log transformation is that a percentage change in the independent variable (for small changes <20%) will illict a constant change in the dependent variable. A square root transformation has a slope which decays asymptotically to zero, so the effect dampens as the values become larger. In the case of density, this represents that increased density in cities may not have as large of an effect. Given the improvement in linearity and normality we will use these transformations moving forward.

```
ggplot(crime_na, aes(x = density, fill = factor(urban))) + geom_histogram() +
    xlab("Density") + ylab("Frequency") + ggtitle("Density Histogram") + scale_fill_discrete(name = "Ur
```

[5] See *Appendix* for the others.



Normality/Skew With and Without Transformation:

```
crime_na["log_prbarr"] = log(crime_na$prbarr)
crime_na["log_prbconv"] = log(crime_na$prbconv)
crime_na["log_polpc"] = log(crime_na$polpc)
crime_na["log_taxpc"] = log(crime_na$taxpc)
crime_na["sqrt_density"] = sqrt(crime_na$density)
```

Finally, we examined collinearity between various variables that we thought would comprise our models. For the sake of narrative flow, discussion of those findings is presented alongside the models in question.

Research Question and Model-Building

Our candidate seeks to develop a policy platform to address crime in NC. However, she knows that both the state's resources and her political capital are limited, so she would like to know how to prioritize her efforts to achieve maximal impact. Our **research question** is thus the following: *What is a small number of factors which are jointly highly predictive of the observable crime rate which our candidate should seek to address through legislation?*

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 additional 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, but we preface our model-building section with this note to as a practical way to invite the reader to critically examine the models we propose.

In order to answer our research question we created several models which included variables related to key crime policy decisions. While there are 25 variables in our raw dataset, our model-building proceeded systematically. First, we grouped variables together *thematically*. Next, we built a model with the variables comprising the theme we believed *a priori* would be most likely to have high predictive value. After evaluating the model's performance, we added the variables for the next theme, and so forth.

Thematic grouping of variables

Theme 1: Crime and Punishment

Our first theme revolves around crime and the likelihood of punishment resulting from the act of committing a crime. We believe that the variables police per capita (`polpc`), probability of arrest (`prbarr`), probability of conviction (`prbconv`), probability of incarceration (`prbpris`), and average sentence length (`avgsen`) fit well together. Taken comprehensively, they help frame the observed crime rate in economic terms as a sort of "risk-reward" proposition for would-be criminals: is the presence of a well-functioning criminal justice system[6]in which crimes consistently lead to arrests (of the actual perpetrators, and not innocent bystanders), and arrests consistently lead to (warranted) convictions, and convictions lead to prison sentences, and sentences are indeed punitive, predictively associated with the crime rate?

The variables in this theme are particularly valuable because they provide a solid foundation for policy development. If the size of the police force is associated with changes in the crime rate, it can be scaled up or scaled down appropriately. If the probability of arrest is impactful, the county could invest in improving community-police relationships with the goal of improving the speed with which crimes are reported and the quality of the intelligence witnesses provide. If the probability of conviction is important, funds could be invested in further training the police and District Attorneys on the collection and presentation of evidence. If the probability of incarceration and average sentence are important, legislators could advocate for changes to sentencing guidelines.

Theme 2: Demography and Economics

Our second theme focuses on issues of community composition, economic resources, and opportunity. Population density (`density`) may influence the crime rate in complex ways: from a sociological perspective, the social ecology approach to studying criminality considers population density to be one of several so-called *criminogenic* ("crime-causing") factors. From a psychological perspective, a larger population may weaken community bonds[7] and thus the pressure to exhibit prosocial behavior. Where significant wealth or income disparities exist within close proximity, relative deprivation theory suggests population density may drive greater criminality. On the other hand, large, sparsely-populated swathes of land are difficult for police to monitor and may present opportunities for various non-violent crimes.[8]

Social control theory suggests that young people with weaker bonds to their communities may be more likely to commit crime. Indeed, economic and criminological research over the last several decades has argued young men are disproportionately involved in crime. There has also been widespread discussion of the role that race and ethnicity play in criminality. We thus include the variables `pctymle` and `pctmin80` in this model.

The police force is funded by the local government, which in turn is funded by the local taxpayers. We might thus expect that the taxes paid per capita (`taxpc`) might have an influence on the size and/or effectiveness of the police force, either reinforcing or undermining any effect `polpc` might have on the crime rate.

[6]Using an admittedly naive formulation in which socioeconomic status, racial bias, and predatory policing are assumed to be non-issues.

[7]Consider the various tiers of Dunbar's number

[8]Consider, for instance, the prevalence of marijuana cultivation in rural northern California or moonshine production in the rural South. Both are industries in which lots of land with very few people living on it is a desirable feature.

The dataset contains nine variables related to wage: `wfed`, `wsta`, `wloc`, `wmfg`, `wser`, `wcon`, `wfir`, `wtrd`, `wtuc`. While much of their cumulative influence on a community may be captured in the `taxpc` variable - and thus we must beware excessive collinearity - different groupings of those wage variables may provide insight into well- or under-funded segments of the local economy. We discuss these alternative specifications below under Model 2.

While we believe these variables are sufficiently distinct to warrant their own model, we believe they may exhibit some collinearity with the predictors we included in **Model 1**. Specifically, we suspect `taxpc` and `density` may be collinear with `polpc`; `pctymle` and `pctmin80` may be collinear with `prbarr`; and government wages (`wfed`, `wsta`, `wloc`) may be collinear with `prbconv`.[9] We theorize that both `prbpris` and `avgsen` may be collinear with some underlying omitted variable (for instance, local sentencing guidelines), but lack an alternative variable that might serve as a reasonable proxy. We will conduct checks for collinearity when analyzing our models.

Theme 3: 'The Kitchen Sink, or, everything else'

Of the 25 variables in the dataset, we incorporated five into **Model 1** and 13 into **Model 2**. Of the remaining seven, `year` is meaningless (it is a constant, 87, for all observations in the dataset) and `crmrte` is our outcome of interest. That leaves five we can incorporate into a final model that attempts to capture everything that might plausibly predict crime rate: three dummy variables describing a county's location within NC (`west`, `central`) and its degree of urbanization (`urban`), one describing the ratio of face-to-face vs. 'other' crimes (`mix`), and the county identifier (`county`), which we know is a FIPS code. The unifying theme for these variables is essentially the absence of a unifying theme: these are the leftovers.

Building of Initial Models

In the section below, we build our models, assess them in light of the six assumptions of the classical linear model (**CLM**), examine their coefficients, and provide some basic interpretation. In addition to the models suggested by the three themes above, we will generate a fourth model which incorporates our insights from evaluating our models.

The CLM Assumptions

There are **six assumptions of the CLM** that we will try to uphold in our models.

1. *Linear population model*: The model is linear in the coefficients of the predictors, correctly specified, and has an additive error term.
2. *Random Sampling*: The data on which the model is based represents a random sample from the population and the generative mechanism produces i.i.d data.
3. *No perfect multicollinearity*: None of the predictors in the models are perfect linear combinations of each other.
4. *Zero-conditional mean / Exogeneity*:

$$E[u|x_i] = 0;\ i = 1...k$$

$$Cov(x_i, u) = 0;\ i = 1...k$$

$$E[u] = 0$$

5. *Homoskedasticity*: Residuals display equal variance for all X_i.

$$Var(u|x_i) = \sigma^2];\ i = 1...k$$

[9] The expected collinearity of government wages with `prbconv` is premised on a theorized relationship between `prbconv` as an outcome variable and the quality or effectiveness of police investigators, court-appointed advocates, judges, and other government officials *as proxied* loosely *by wage*.

6. *Normality of Errors*: Residuals will be normally distributed.

We cannot empirically test **CLM2**, but we will assume it is met as our "universe" is counties in N.C and we have data on the entire population.

Model 1: Crime and Punishment

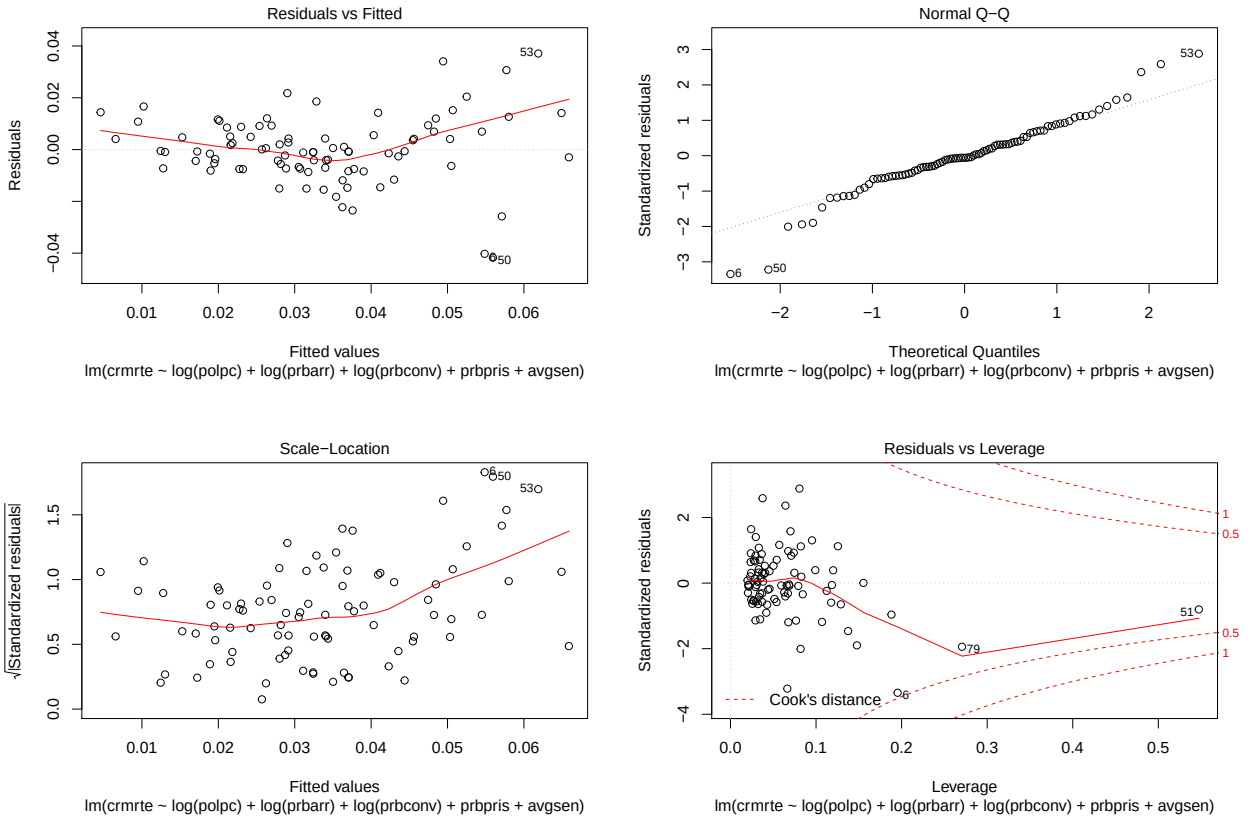
Drawing upon the transformations suggested by our exploratory data analysis and the discussion above, the first model we develop and examine is thus as follows:

$$\texttt{crmrte} = \beta_0 + \beta_1 \log(\texttt{polpc}) + \beta_2 \log(\texttt{prbarr}) + \beta_3 \log(\texttt{prbconv}) + \beta_4 \texttt{prbpris} + \beta_5 \texttt{avgsen} + u \quad (1)$$

```
# Model for Theme 1
mod1 <- 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.
mod1$AIC <- AIC(mod1)
```

Before we examine the coefficients of our model, we generate our diagnostic plots to see whether the model adheres to the six assumptions of the classical linear model.

```
# par(mfrow = c(3, 2))
plot(mod1)
```



While the mean of the residuals appears to be centered on 0 as we would expect (supporting **CLM4**), the curvature in our loess smoother suggests that our model may not be correctly specified (violating **CLM1**). Our residuals appear to be quite normal (supporting **CLM6**), except at the extremes of our model where few data points are available. We can confirm or refute that interpretation using the Shapiro-Wilke test for normality and applying it to the residuals.

```
shapiro.test(mod1$residuals) %>% tidy() %>% kable()
```

statistic	p.value	method
0.9629367	0.0114959	Shapiro-Wilk normality test

The p-value of 0.011 indicates that our residuals are not fully normal, however, significance is not an unexpected result with our somewhat large dataset.

We appear to see some evidence of heteroskedasticity (violating **CLM5**), indicating a need to apply heteroskedasticity-consistent ("robust") standard errors.

```
bptest(mod1) %>% tidy() %>% kable()  # Breusch-Pagan: heteroskedasticity confirmed
```

statistic	p.value	parameter	method
22.11243	0.0004984	5	studentized Breusch-Pagan test

```
ncvTest(mod1)  # Non-constant variance: heteroskedasticity confirmed
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 32.97464    Df = 1     p = 9.336874e-09
```

Having confirmed the heteroskedasticity in our model, we will need to replace the standard errors and the F-statistic with robust versions of themselves.

Before we do, however, we can check for multicollinearity in our model by examining the variance inflation factors (VIF) for each of the predictors we include; a VIF above 5 for any of them would represent cause for concern, and indicate that we could likely eliminate that predictor without significant loss of performance by the model.

```
vif(mod1) %>% t() %>% kable()  # All clear
```

log(polpc)	log(prbarr)	log(prbconv)	prbpris	avgsen
1.284925	1.047115	1.084671	1.012952	1.280973

Having found no indication that we should remove existing predictors from our model, we compute our robust standard errors and F-statistic to add to our reporting. We demonstrate the code here, but will omit it from our discussion of future models.[10]

```
mod1_cov <- vcovHC(mod1, type = "HC1")
mod1_pvals <- list(coeftest(mod1, vcov = mod1_cov)[, 4])
mod1_robse <- list(sqrt(diag(mod1_cov)))
mod1_wald <- waldtest(mod1, vcov = mod1_cov)
mod1_wDf <- mod1_wald$Df[2]
mod1_wF <- round(mod1_wald$F[2], 2)
```

[10] See *Appendix* for detailed code

```
mod1_wp <- formatC(mod1_wald$`Pr(>F)`[2], format = "e", digits = 2)
mod1_summ <- summary.lm(mod1)
```

```
# Output model results in nice format using tidy and kable kable(tidy(mod1))
# # <- TS: Superfluous b/c need RSE, so stargazer
stargazer(mod1, header = FALSE, type = "latex", se = mod1_robse, p = mod1_pvals,
    report = "vc*s", star.cutoffs = c(0.05, 0.01, 0.001), title = "Model 1: Crime and Punishment",
    single.row = TRUE, omit.stat = "f", add.lines = list(c("F Statistic", mod1_wF),
        c("Pr(>F)", mod1_wp)), notes = "p-values based on Wald test; robust standard errors in parenthes
```

Table 6: Model 1: Crime and Punishment

	Dependent variable:
	crmrte
log(polpc)	0.020^{***} (0.005)
log(prbarr)	-0.024^{***} (0.004)
log(prbconv)	-0.014^{***} (0.004)
prbpris	0.009 (0.021)
avgsen	-0.001 (0.001)
Constant	0.126^{***} (0.036)
F Statistic	11.09
Pr(>F)	3.22e-08
Observations	90
R^2	0.524
Adjusted R^2	0.495
Akaike Inf. Crit.	-512.807
Residual Std. Error	0.013 (df = 84)
Note:	*p<0.05; **p<0.01; ***p<0.001
	p-values based on Wald test; robust standard errors in parentheses

Comments on Model 1: Crime and Punishment

The p-value associated with our heteroskedasticity-robust F-test is highly significant, indicating that our model is more predictive than a simple horizontal line at a y-intercept would be. The adjusted R^2 of our model is 0.495, which is reasonably high given the complexity of crime.

We highlight four main observations:

1) The log of police per capita, the log of the probability of arrest, and the log of probability of conviction are all highly statistically-significant. Because these predictors are all log transformed while the outcome variable is not, the interpretation is that the coefficient represents the change in the predicted value of the outcome variable associated with a 1% change in the predictor. In other words, $\Delta y = \Delta x_i/100$.

2) Our coefficient for (log(*polpc*)) is positive, and suggests that a 1% increase in the # of police per capita would be associated with an *increase* of 2.05 crimes per 10,000 people. 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.

3) Both log probability of arrest (log(*prbarr*)) and log probability of conviction (log(*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. Our model suggests a 1% increase

in the 'probability' of arrest and conviction would be associated with -2.44 and -1.37 fewer crimes per 10,000 people.

4) Neither the probability of incarceration of the average sentence length are statistically significant.

Model 2: Demography and Economics

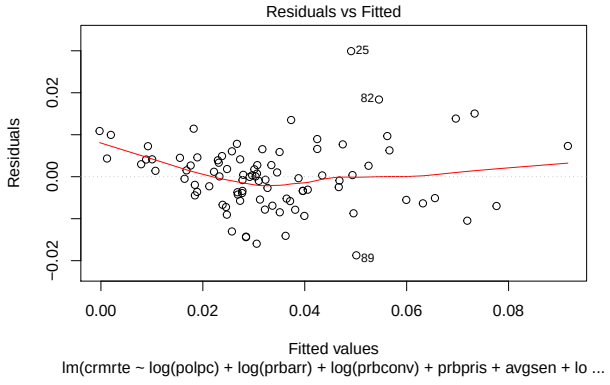
Model 2 builds upon **Model 1** by including the additional covariates discussed earlier. If we are correct in our assumption that these are in fact relevant covariates, our model fit should improve and we should help mitigate omitted variable bias. The base (extended) version of **Model 2** would be as follows:

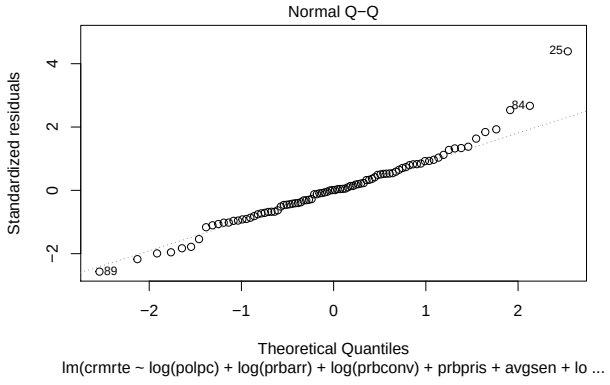
$$\begin{aligned} \texttt{crmrte} = \beta_0 &+ \beta_1 \log(\texttt{polpc}) + \beta_2 \log(\texttt{prbarr}) + \beta_3 \log(\texttt{prbconv}) + \beta_4 \texttt{prbpris} + \beta_5 \texttt{avgsen} \qquad \text{(Model 1)} \\ &+ \beta_6 \log(\texttt{taxpc}) + \beta_7 \sqrt{\texttt{density}} + \beta_8 \texttt{pctymle} + \beta_9 \texttt{pctmin80} \\ &+ \beta_{10} \texttt{wcon} + \beta_{11} \texttt{wtuc} + \beta_{12} \texttt{wtrd} + \beta_{13} \texttt{wfir} + \beta_{14} \texttt{wser} \\ &+ \beta_{15} \texttt{wmfg} + \beta_{16} \texttt{wfed} + \beta_{17} \texttt{wsta} + \beta_{18} \texttt{wloc} + u \qquad (2) \end{aligned}$$

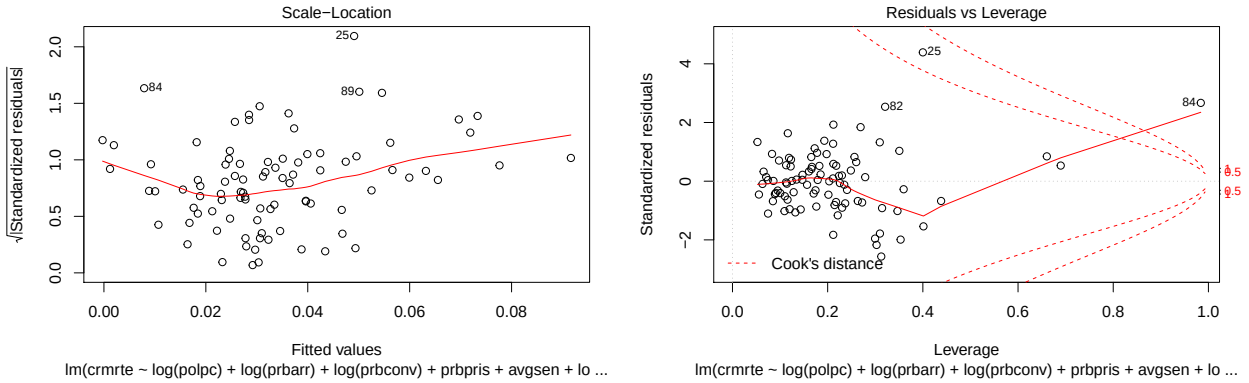
```
# Model with our key explanatory variables, and what we suspect to be key
# covariates
mod2_b <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    prbpris + avgsen + log(taxpc) + sqrt(density) + pctymle + pctmin80 + wcon +
    wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc))

# Adding AIC to our model to help us compare models in the future.
mod2_b$AIC <- AIC(mod2_b)
```

Before we examine the coefficients of our model, we generate our diagnostic plots to see whether the model adheres to the six assumptions of the classical linear model.

```
# par(mfrow = c(3, 2))
plot(mod2_b)
```







Overall, the most salient feature of these diagnostic plots is how similar they are to the plots for **Model 1**. That being said, we do see improvement in the plot of residuals vs. fitted values as there appears to be less correlation between our independent variable and the residuals. Additionally, while there may be some non-linearity (suggesting a violation of zero conditional mean) it occurs at very low and high predicted values, where we have very few data points, and therefore we will operate under the assumption that our zero conditional mean assumption is valid.

One interesting difference is that two observations - for the counties at indices 25 and 84 - are now both high-leverage and high-influence, with Cook's distances of > 0.5 and > 1.0 respectively. Which counties might those represent?

```
kable(crime_na[c(25, 84), c(1:10)])
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc
55	87	0.0790163	0.224628	0.207831	0.304348	13.57	0.0040096	0.5115089	119.76145
185	87	0.0108703	0.195266	2.121210	0.442857	5.38	0.0012221	0.3887588	40.82454

```
kable(crime_na[c(25, 84), c(14:23)])
```

pctmin80	wcon	wtuc	wtrd	wfir	wser	wmfg	wfed	wsta	wloc
6.49622	309.5238	445.2762	189.7436	284.5933	221.3903	319.21	338.91	361.68	326.08
64.34820	226.8245	331.5650	167.3726	264.4231	2177.0681	247.72	381.33	367.25	300.13

```
kable(crime_na[c(25, 84), c(25)])
```

pctymle
0.0761381
0.0700822

Interestingly, neither of those is county 115, which had been the county we flagged as having outliers in `crmrte`, `prbarr`, and `polpc`. However, we can see that county 185 is the one with the exceedingly high `wser`.

While the plot suggests that our residuals again deviate from normality at the extremes, this time, the Shapiro-Wilke test of our residuals ($p = 0.119$) suggests that they are normally distributed (supporting **CLM6**).

```r
vif(mod2_b) %>% kable(col.names = c("VIF"))  # All clear
```

	VIF
log(polpc)	2.099671
log(prbarr)	1.730140
log(prbconv)	1.866910
prbpris	1.154503
avgsen	1.556323
log(taxpc)	1.745862
sqrt(density)	3.014007
pctymle	1.466395
pctmin80	1.297078
wcon	2.135526
wtuc	1.720212
wtrd	3.143922
wfir	2.650859
wser	1.239784
wmfg	1.947060
wfed	3.076304
wsta	1.510677
wloc	2.235522

Our analysis of the VIFs indicates that multicollinearity has increased slightly (median = 1.81, mean = 1.98 max = 3.14), but not yet to the point where our rule of thumb (excluding covariates with VIF > 5) would come into play.

As with **Model 1**, graphical indications of heteroskedasticity (violating **CLM5**) are confirmed by Breusch-Pagan ($p \approx 0.00074$) and non-constant variance ($p \approx 0.00226$) diagnostic tests, indicating a need to apply heteroskedasticity-consistent ("robust") standard errors.[11]

In summary, model 2 satisfies the Classical Linear Model Assumptions of linearity (CLM1), no perfect multicollinearity (CLM3), zero conditional mean (CLM4), and normality of errors (CLM6). While our homoskedasticity assumption is violated (CLM5) we are fortunately able to correct for this by using heteroskedasticity robust standard errors.

Comments on Model 2: Demography and Economics

There are a few observations to make with respect to **Model 2**.

1) First, all of the predictors that were significant under **Model 1** remained so under this expanded model. While the magnitude of their coefficients has uniformly decreased, the signs (indicating the direction of the relationship) have remained the same.

2) 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.

```r
mod2_b_nomin <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    prbpris + avgsen + log(taxpc) + sqrt(density) + pctymle + wcon + wtuc +
    wtrd + wfir + wser + wmfg + wfed + wsta + wloc))

min_ars_diff <- summary(mod2_b)$adj.r.squared - summary(mod2_b_nomin)$adj.r.squared
```

[11] See *Appendix* for the detailed code showing the calculation of those SEs

Table 11: Model 2: Demography and Economics

	Dependent variable:
	crmrte
log(polpc)	0.014^{**} (0.005)
log(prbarr)	-0.016^{***} (0.004)
log(prbconv)	-0.008^{*} (0.003)
prbpris	-0.010 (0.011)
avgsen	-0.001 (0.0004)
log(taxpc)	0.006 (0.007)
sqrt(density)	0.018^{***} (0.004)
pctymle	0.064 (0.033)
pctmin80	0.0004^{***} (0.0001)
wcon	0.00002 (0.00002)
wtuc	0.00001 (0.00002)
wtrd	0.00004 (0.0001)
wfir	-0.00005 (0.00002)
wser	-0.00001^{***} (0.00000)
wmfg	-0.00000 (0.00001)
wfed	0.00000 (0.00003)
wsta	-0.00001 (0.00003)
wloc	0.00003 (0.0001)
Constant	0.047 (0.048)
F Statistic	19.27
Pr(>F)	1.79e-20
Observations	90
R^2	0.827
Adjusted R^2	0.783
Akaike Inf. Crit.	-577.971
Residual Std. Error	0.009 (df = 71)

Note: *p<0.05; **p<0.01; ***p<0.001
p-values based on Wald test; robust standard errors in parentheses

When including the minority variable in our model, the adjusted R^2 is 0.783. When excluding it, the adjusted R^2 is 0.661, a difference of 12.3 percentage points.

3) The square root of the population density is also highly significant. This suggests that the observed crime rate grows quickly as the population density changes from very low (sparsely populated) to medium density, but beyond a certain point each additional person per square mile should influence our prediction of the crime rate less and less.

4) In this model, the wages associated with the service industry (`wser`) have become significant. This is somewhat unexpected - why should a single industry's wages matter, when others do not? We recall that the service industry in county 185 had a weekly average wage nearly 10 times that of the other industries - and also that our diagnostic plots showed that it had a Cook's distance > 1. We can re-run this model with a dataset that excludes county 185 to see if the finding holds when the outlier is removed.

```
mod2_bnout <- lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) + prbpris +
    avgsen + log(taxpc) + sqrt(density) + pctymle + pctmin80 + wcon + wtuc +
    wtrd + wfir + wser + wmfg + wfed + wsta + wloc, data = subset(crime_na,
    county != max(county)))
mod2_bnout$AIC <- AIC(mod2_bnout)
```

```
stargazer(mod2_b, mod2_bnout, header = FALSE, type = "latex", column.labels = c("All Obs",
    "Excl. 185"), se = c(mod2_b_robse, mod2_bnout_robse), p = c(mod2_b_pvals,
    mod2_bnout_pvals), report = "vc*s", star.cutoffs = c(0.05, 0.01, 0.001),
    title = "Model 2: Demography and Economics", single.row = TRUE, omit.stat = "f",
    add.lines = list(c("F Statistic", mod2_b_wF, mod2_bnout_wF), c("Pr(>F)",
        mod2_b_wp, mod2_bnout_wp)), notes = "p-values based on Wald test; robust standard errors in par
```

Table 12: Model 2: Demography and Economics

	Dependent variable:	
	crmrte	
	All Obs	Excl. 185
	(1)	(2)
log(polpc)	0.014** (0.005)	0.014** (0.005)
log(prbarr)	−0.016*** (0.004)	−0.016*** (0.004)
log(prbconv)	−0.008* (0.003)	−0.008* (0.003)
prbpris	−0.010 (0.011)	−0.010 (0.011)
avgsen	−0.001 (0.0004)	−0.001 (0.0004)
log(taxpc)	0.006 (0.007)	0.006 (0.007)
sqrt(density)	0.018*** (0.004)	0.018*** (0.004)
pctymle	0.064 (0.033)	0.067* (0.033)
pctmin80	0.0004*** (0.0001)	0.0004*** (0.0001)
wcon	0.00002 (0.00002)	0.00001 (0.00002)
wtuc	0.00001 (0.00002)	0.00001 (0.00002)
wtrd	0.00004 (0.0001)	0.00005 (0.0001)
wfir	−0.00005 (0.00002)	−0.00005* (0.00002)
wser	−0.00001*** (0.00000)	−0.00001*** (0.00000)
wmfg	−0.00000 (0.00001)	−0.00000 (0.00001)
wfed	0.00000 (0.00003)	0.00000 (0.00003)
wsta	−0.00001 (0.00003)	−0.00001 (0.00003)
wloc	0.00003 (0.0001)	0.00003 (0.0001)
Constant	0.047 (0.048)	0.045 (0.048)
F Statistic	19.27	18.77
Pr(>F)	1.79e-20	5.49e-20
Observations	90	89
R^2	0.827	0.826
Adjusted R^2	0.783	0.781
Akaike Inf. Crit.	−577.971	−570.451
Residual Std. Error	0.009 (df = 71)	0.009 (df = 70)

Note: *p<0.05; **p<0.01; ***p<0.001
p-values based on Wald test; robust standard errors in parentheses

Indeed, we see that it does. Practically speaking, however, it is unclear whether the magnitude of the effect is significant. It suggests that a one-unit (i.e., one dollar) increase in the average weekly service-sector wage is associated with a change of -0.12 (i.e., decrease) in the number of crimes per 10,000 people, which in turn implies that an additional $10/week would be associated with one fewer crime. While $10 over the course of the week doesn't sound like much, the average across all counties is only $275.34; such an increase would be the equivalent of a 3.63% raise across the entire sector.

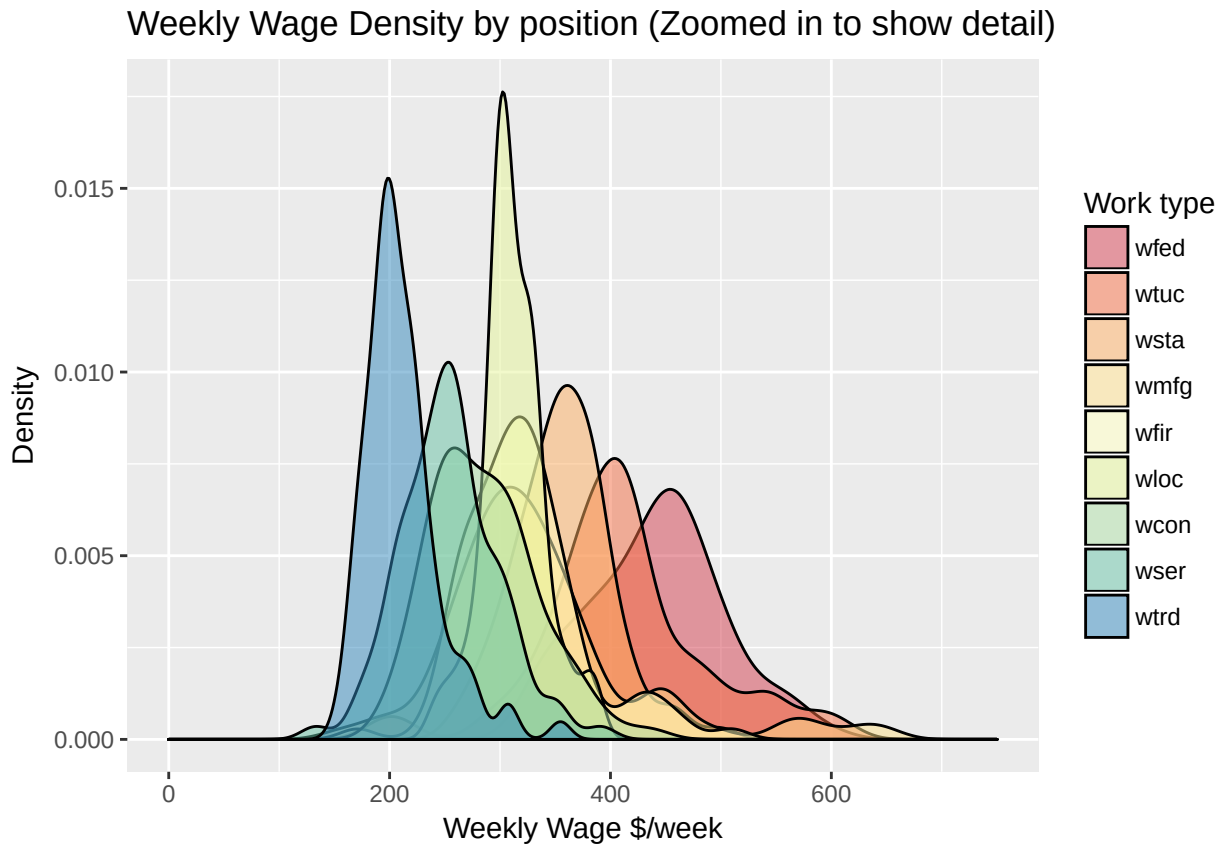
Applying our interpretive lens to $\sqrt{\texttt{density}}$ is not so straightforward. This is one of the tradeoffs we make for improving our model fit: we may sacrifice the ability to easily explain the effect. So we will not attempt to explain it beyond the earlier observation that at higher population densities, each marginal person per square mile modifies our prediction of crime rate less and less.

The `pctmin80` interpretation is more straightforward, although again it is important to recall that our model is merely predictive and does not provide causal explanations. Each increase in the percentage of minorities in the population in 1980 is associated with 4.04 more crimes per 10,000 people.

Joint Wage Significance

In the previous section we discussed the significance of the "service wage" variable. A single wage variable might be significant just by chance when there are 9 such variables, so one way to gauge if wage levels in general are predictive would be to calculate an F statistic for a joint hypothesis test. First, we will look at the distributions of wage variables to ensure nothing is too abnormal:

```
wage_plot <- 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 (Zoomed in to show detail)") +
    xlab("Weekly Wage $/week") + ylab("Density") + labs(fill = "Work type") +
    xlim(0, 750)
wage_plot
```



None of the distributions look too extreme, and therefore we will calculate the F statistic for our joint hypothesis (All wage coefficients equal to zero):

$$F = \frac{(SSR_{\text{restricted}} - SSR_{\text{unrestricted}})/q}{SSR_{\text{unrestricted}}/(n-k-1)}$$

```
# Conduct join hypothesis F-test to determine if wages are significant when
# considered collectively

linearHypothesis(mod2_b, c("wcon=0", "wtuc=0", "wtrd=0", "wfir=0", "wser=0",
    "wmfg=0", "wfed=0", "wsta=0", "wloc+0"), white.adjust = "hc1")
```

```
## Linear hypothesis test
##
## Hypothesis:
## wcon = 0
## wtuc = 0
## wtrd = 0
## wfir = 0
## wser = 0
## wmfg = 0
## wfed = 0
## wsta = 0
## wloc = 0
##
## Model 1: restricted model
```

```
## Model 2: crmrte ~ log(polpc) + log(prbarr) + log(prbconv) + prbpris +
##     avgsen + log(taxpc) + sqrt(density) + pctymle + pctmin80 +
##     wcon + wtuc + wtrd + wfir + wser + wmfg + wfed + wsta + wloc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F  Pr(>F)
## 1     80
## 2     71  9 2.0396 0.04708 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When considered collectively wage levels are significant at the .05 significance level.

Model 3

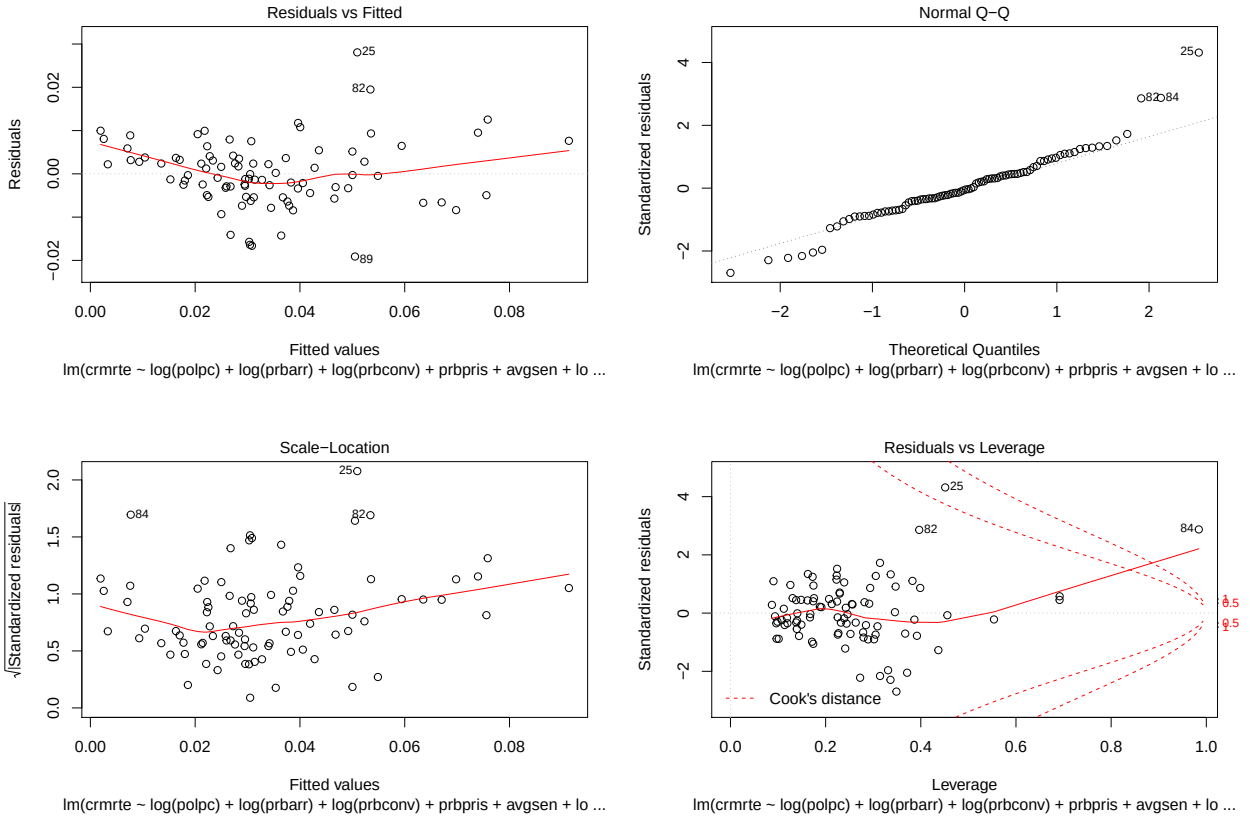
Model 2 yielded greater predictive value than **Model 1** by including a few extra covariates. **Model 3** throws in every available variable in the dataset.

$$\begin{aligned}
\texttt{crmrte} = \beta_0 &+ \beta_1 \log(\texttt{polpc}) + \beta_2 \log(\texttt{prbarr}) + \beta_3 \texttt{prbconv} + \beta_4 \texttt{prbpris} + \beta_5 \texttt{avgsen} && \text{(Model 1)} \\
&+ \beta_6 \log(\texttt{taxpc}) + \beta_7 \sqrt{\texttt{density}} + \beta_8 \texttt{pctymle} + \beta_9 \texttt{pctmin80} \\
&+ \beta_{10} \texttt{wcon} + \beta_{11} \texttt{wtuc} + \beta_{12} \texttt{wtrd} + \beta_{13} \texttt{wfir} + \beta_{14} \texttt{wser} \\
&+ \beta_{15} \texttt{wmfg} + \beta_{16} \texttt{wfed} + \beta_{17} \texttt{wsta} + \beta_{18} \texttt{wloc} && \text{(Model 2)} \\
&+ \beta_{19} \texttt{mix} + \beta_{20} \texttt{central} + \beta_{21} \texttt{west} + \beta_{22} \texttt{urban} + \beta_{23} \texttt{county} + u && (3)
\end{aligned}$$

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.

```
# Linear model including our key explanatory variables, suspected
# covariates, and most other variables
mod3 <- with(crime_na, lm(crmrte ~ log(polpc) + log(prbarr) + log(prbconv) +
    prbpris + avgsen + log(taxpc) + sqrt(density) + pctymle + pctmin80 + west +
    central + urban + mix + wcon + wtuc + wtrd + wfir + wser + wmfg + wfed +
    wsta + wloc))

# Adding AIC to our model to help us compare models in the future.
mod3$AIC <- AIC(mod3)
```

```
plot(mod3)
```



Similar to the other models, the mean of the residuals appears to be centered on 0. Based on the Normal QQ plot our residuals appear to be fairly normal, except at the extremes of our model where there are fewer data points. Again, we apply the Shapiro-Wilke test for normality to our **Model 3** residuals.

```
##
##  Shapiro-Wilk normality test
##
## data:  mod3$residuals
## W = 0.97072, p-value = 0.03977
```

Since the test results in a p value of less than 0.05 we reject the null hypothesis that the underlying distribution of the residuals in normal. To compensate, we will use heteroskedasticity robust statistical versions of the standard error, F and other values.

```
mod3_cov <- vcovHC(mod3, type = "HC1")
mod3_pvals <- list(coeftest(mod3, vcov = mod3_cov)[, 4])
mod3_robse <- list(sqrt(diag(mod3_cov)))
mod3_wald <- waldtest(mod3, vcov = mod3_cov)
mod3_wDf <- mod3_wald$Df[2]
mod3_wF <- round(mod3_wald$F[2], 2)
mod3_wp <- formatC(mod3_wald$`Pr(>F)`[2], format = "e", digits = 2)
mod3_summ <- summary.lm(mod3)
```

Table 13: Model 3: The Kitchen Sink

	Dependent variable:
	crmrte
log(polpc)	0.016^{**} (0.005)
log(prbarr)	-0.014^{***} (0.003)
log(prbconv)	-0.008^{*} (0.003)
prbpris	−0.005 (0.011)
avgsen	−0.001 (0.0004)
log(taxpc)	0.003 (0.007)
sqrt(density)	0.017^{***} (0.004)
pctymle	0.044 (0.038)
pctmin80	0.0003^{***} (0.0001)
west	−0.004 (0.004)
central	−0.005 (0.003)
urban	0.004 (0.006)
mix	−0.015 (0.020)
wcon	0.00002 (0.00003)
wtuc	0.00001 (0.00002)
wtrd	0.00005 (0.0001)
wfir	−0.00004 (0.00002)
wser	-0.00001^{*} (0.00000)
wmfg	−0.00001 (0.00001)
wfed	−0.00000 (0.00003)
wsta	−0.00001 (0.00003)
wloc	0.00002 (0.0001)
Constant	0.076 (0.053)
F Statistic	21.44
Pr(>F)	2.84e-22
Observations	90
R^2	0.837
Adjusted R^2	0.784
Akaike Inf. Crit.	−575.489
Residual Std. Error	0.009 (df = 67)

Note: *p<0.05; **p<0.01; ***p<0.001
p-values based on Wald test; robust standard errors in parentheses

Comments on Model 3: The Kitchen Sink

Similar 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 0.784 which represents a negligible improvement over our previous model.

Model Comparison

Below is a comparison table for our three initial models. 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. This behavior is expected because AIC is a measure of explanatory power and parsimony, where a model with a larger number of variables is penalized. This is an attractive property of AIC because it illustrates how a model can become over fitted and brittle when an excess of independent variables are included.

```
stargazer(mod1, mod2_b, mod3, type = "latex", report = "vc*s", header = FALSE,
    title = "Linear Models Predicting Crime Rate", single.row = TRUE, 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 omitted variable bias, which can invalidate our ability to assume each explanatory variable is uncorrelated with the error term. While we may not always be able mitigate this by including the variable, there are times when we can calculate the direction of the bias.

As a quick refresher to the reader, the Bias in the coefficient of X1 as a result of an omitted variable X2 is equal to the product of the omitted variable's coefficient if it were to be included in the population model (Beta 2), and the coefficient belonging to X1 when X2 is regressed on X1 (delta).

True Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Omitted variable as function of included variables:

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

Bias

$$Bias_{\beta_1} = \delta_1 \beta_2$$

While there are infinite variables which exist, there are several which we believe deserve commentary:

1) Political Party in Control: Traditionally crime policy and legislation is a 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"), our assumption would be a positive correlation between "is_repulican" and the two aforementioned variables. Unfortunately, there isn't

Table 14: Linear Models Predicting Crime Rate

	Dependent variable:		
	crmrte		
	(1)	(2)	(3)
log(polpc)	0.020*** (0.004)	0.014*** (0.004)	0.016*** (0.004)
log(prbarr)	−0.024*** (0.004)	−0.016*** (0.003)	−0.014*** (0.003)
log(prbconv)	−0.014*** (0.003)	−0.008*** (0.002)	−0.008*** (0.002)
prbpris	0.009 (0.018)	−0.010 (0.012)	−0.005 (0.013)
avgsen	−0.001 (0.001)	−0.001 (0.0004)	−0.001* (0.0004)
log(taxpc)		0.006 (0.005)	0.003 (0.005)
sqrt(density)		0.018*** (0.003)	0.017*** (0.004)
pctymle		0.064 (0.048)	0.044 (0.050)
pctmin80		0.0004*** (0.0001)	0.0003*** (0.0001)
west			−0.004 (0.004)
central			−0.005 (0.003)
urban			0.004 (0.006)
mix			−0.015 (0.016)
wcon		0.00002 (0.00003)	0.00002 (0.00003)
wtuc		0.00001 (0.00002)	0.00001 (0.00002)
wtrd		0.00004 (0.00005)	0.00005 (0.00005)
wfir		−0.00005 (0.00003)	−0.00004 (0.00003)
wser		−0.00001** (0.00001)	−0.00001* (0.00001)
wmfg		−0.00000 (0.00001)	−0.00001 (0.00002)
wfed		0.00000 (0.00003)	−0.00000 (0.00003)
wsta		−0.00001 (0.00003)	−0.00001 (0.00003)
wloc		0.00003 (0.00005)	0.00002 (0.0001)
Constant	0.126*** (0.033)	0.047 (0.042)	0.076 (0.048)
Observations	90	90	90
R^2	0.524	0.827	0.837
Adjusted R^2	0.495	0.783	0.784
Akaike Inf. Crit.	−512.807	−577.971	−575.489

any theory to guide us on whether "is_republican" would be correlated with crime rate. As such we can't accurately anticipate the direction of the bias. While outside the scope of this research we could either a) perform an ancillary exploration to determine the correlation between crime rates and political party, or b) append political party to our data.

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 positively correlated with higher rates of crime as people who may not normally commit criminal activity are pushed to their limits. We would anticipate that tax revenue per capita is negatively correlated with unemployment, which would in turn, suggest a negative bias. We might also wish to be more granular, and include both minority and majority unemployment rates to help control for racial inequality. This data could be obtained from the Bureau of Labor Statistics; we leave this next step to future researchers.

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 phenomenon exists. Because this is a phenomenon which occurs mainly in cities, we would expect our normalized standard deviation variable to be positively correlated with both density and crime rate. As such, we would anticipate there to be a positive bias in the coefficient for density.

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. We suspect that a "guardian" mindset is negatively correlated with arrest and conviction rates, but unfortunately, we have no a priori beliefs regarding the correlation between crime rates and police methodology prohibiting our ability to predict bias direction.

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. Like 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. If we express police representation as a variable from 0 to 1, with 1 being perfect representation, then we would expect this variable to be negatively correlated with crime rates (we presume this indicates a closer police community bond). We might also reason that it would be negatively correlated with arrest rates, suggesting our coefficient for arrest rates would have a positive bias.

6) Criminals' Perception of Risk-Reward Ratios: Of particular relevance to the Model 1 ("Crime and Punishment") analysis is the issue of criminals' perception of the "effectiveness'"of the justice system. The model makes two critical assumptions: first, that criminality is driven by rational behavior; second, that criminals have sufficient insight into the outcomes of criminal justice proceedings for their decisions (which we are assuming to be rational) to be informed by them. Obviously if criminals are not of the genus *Homo economicus*, a major theoretical underpinning of the model is removed. However, even if criminals *are* rational actors, they may be operating with imperfect information: they may make the "incorrect" risk-reward calculation if they are unaware of how risky their criminality truly is. For instance, a thief who does not know the District Attorney always 'gets her man' may not be deterred by a high `prbconv` as our model might assume. In a perfect model we would have variable which captures the degree to which individuals both have accurate information and act rationally. We might assume that higher levels of rationality are negatively correlated with crime rates as well as with percentage of young males; therefore we would expect the coefficient for our "pctymle" variable to have a positive bias.

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 recommendations 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

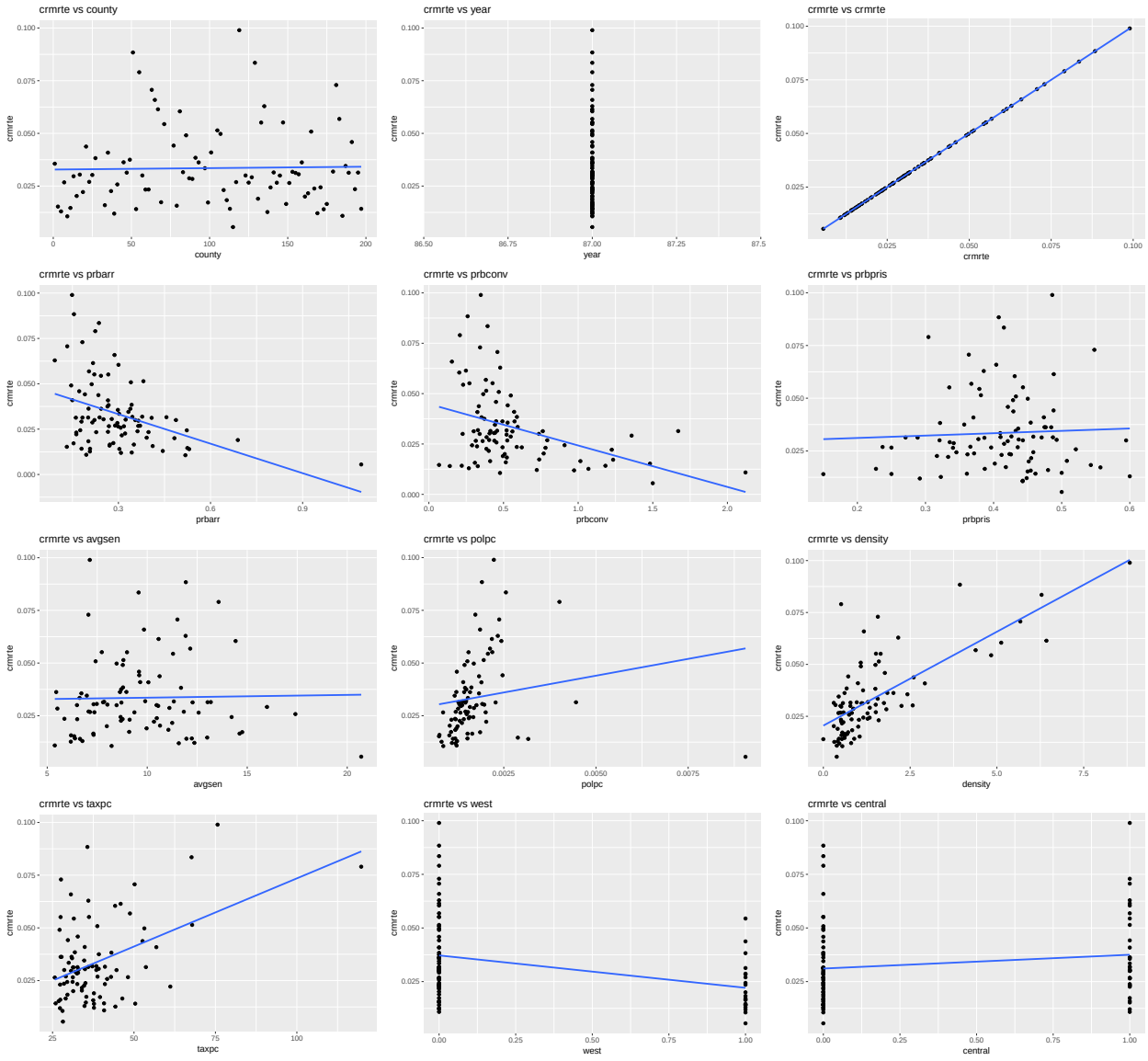
The relationships suggested by our models are intuitive: an increase in the number of police per capita and the population density in an area are associated with (predictive of, in our model) increases in the crime rate; increases in the probability of arrest and probability of conviction are associated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

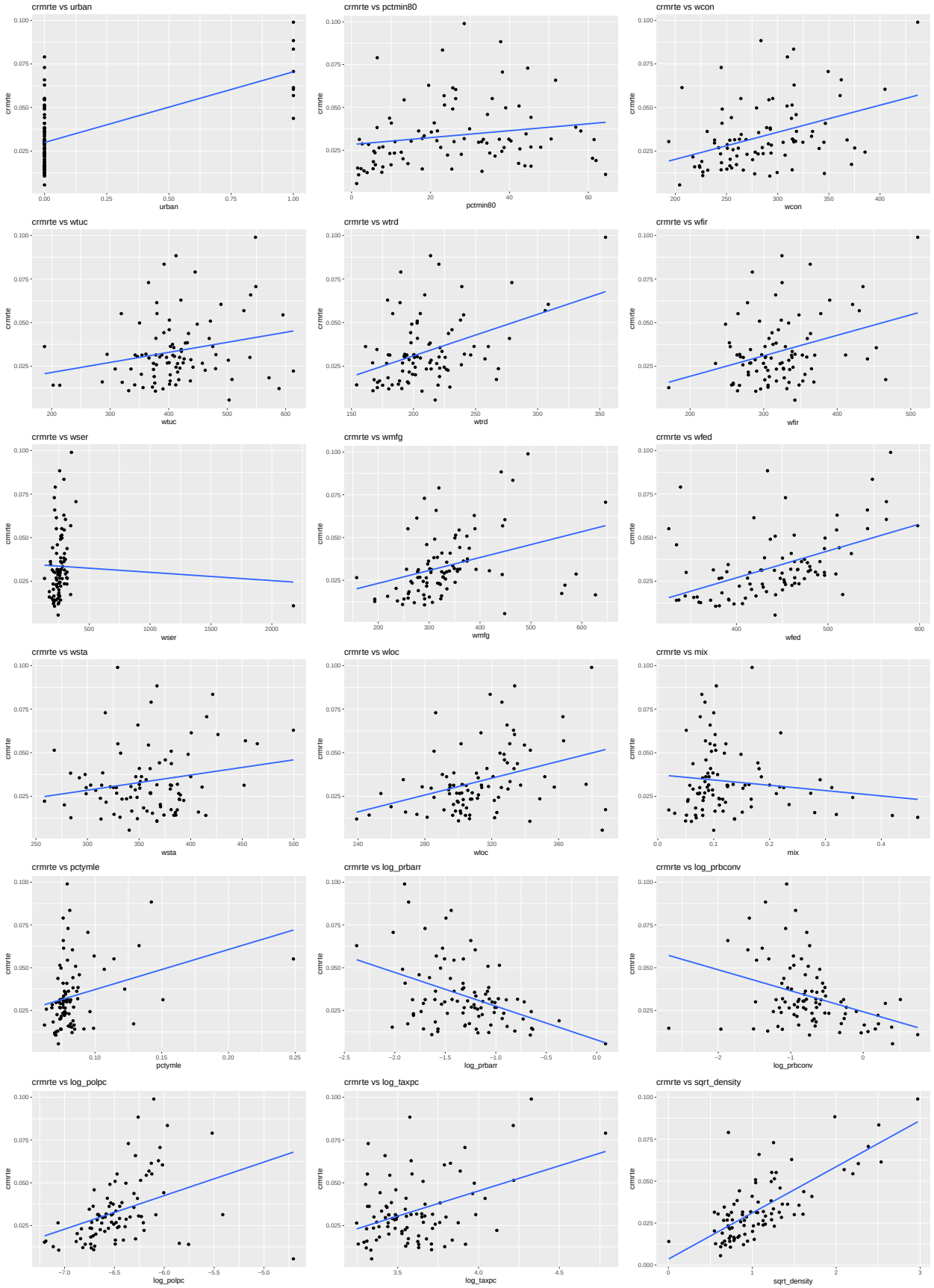
Four of them are simple to interpret: a 1% increase in the number of police is associated with an increase in the crime rate of 11.1 crimes per 100,000 people. Increasing by 1% the probability that a crime report leads to an arrest is associated with a decrease in the crime rate of 15.9 crimes per 100,000 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

Appendix

Detail: scatterplots of variables generated during EDA.

```
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(substitute(y)) + ggtitle(str_glue("{y} vs {v}")))
    }
}
make_scatters(crime_na, var_list = names(crime_na), "crmrte")
```




crmrte vs urban
crmrte vs pctmin80
crmrte vs wcon
crmrte vs wtuc
crmrte vs wtrd
crmrte vs wfir
crmrte vs wser
crmrte vs wmfg
crmrte vs wfed
crmrte vs wsta
crmrte vs wloc
crmrte vs mix
crmrte vs pctymle
crmrte vs log_prbarr
crmrte vs log_prbconv
crmrte vs log_polpc
crmrte vs log_taxpc
crmrte vs sqrt_density
crmrte
urban
pctmin80
wcon
wtuc
wtrd
wfir
wser
wmfg
wfed
wsta
wloc
mix
pctymle
log_prbarr
log_prbconv
log_polpc
log_taxpc
sqrt_density

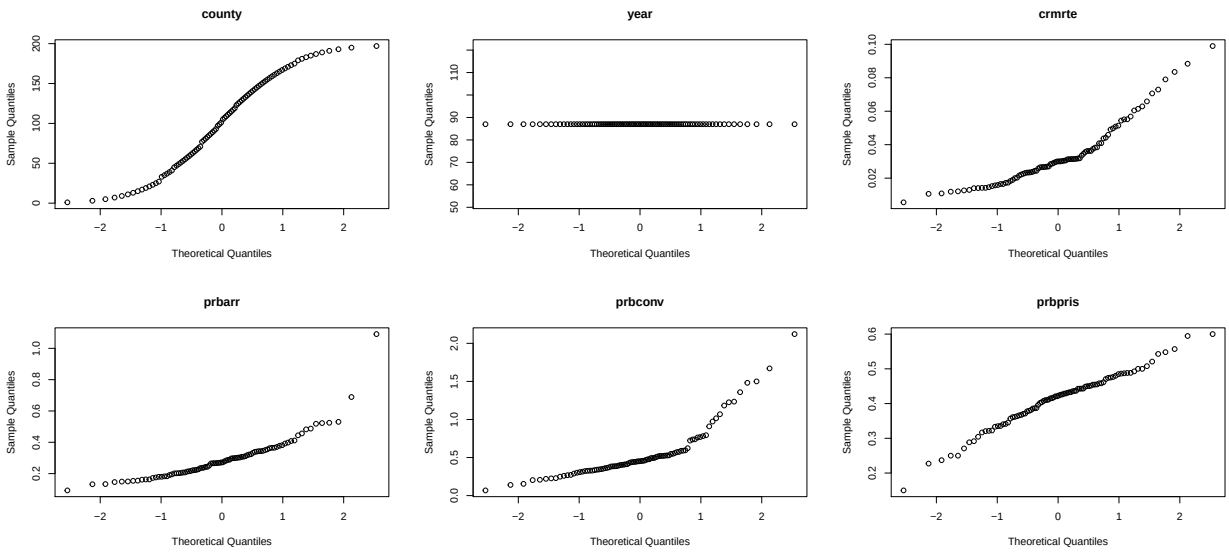
Detail: observations in which the 'probability' variables did not behave as probabilities insofar as they fell outside the range of 0:1.

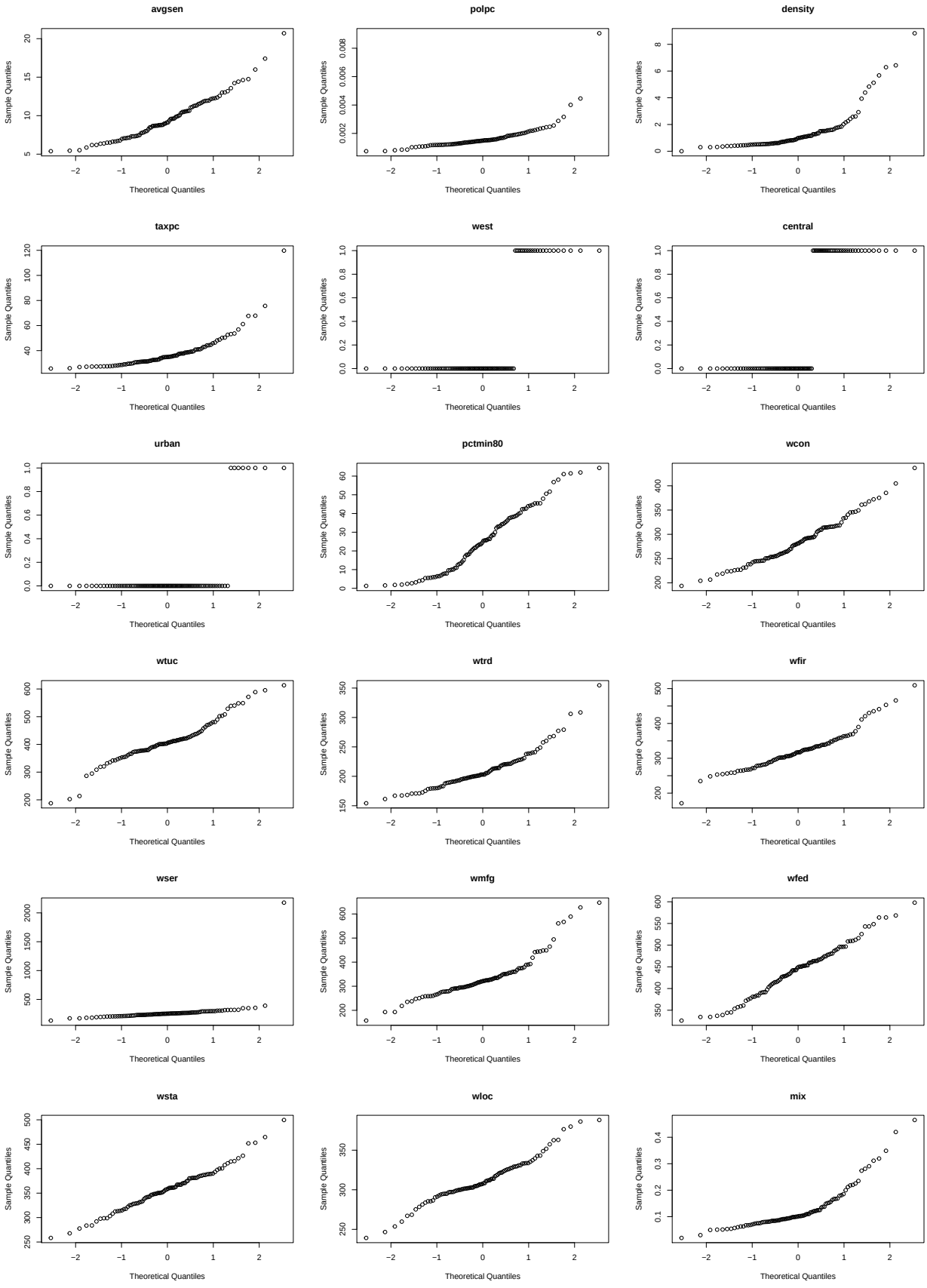
```
non_prob
```

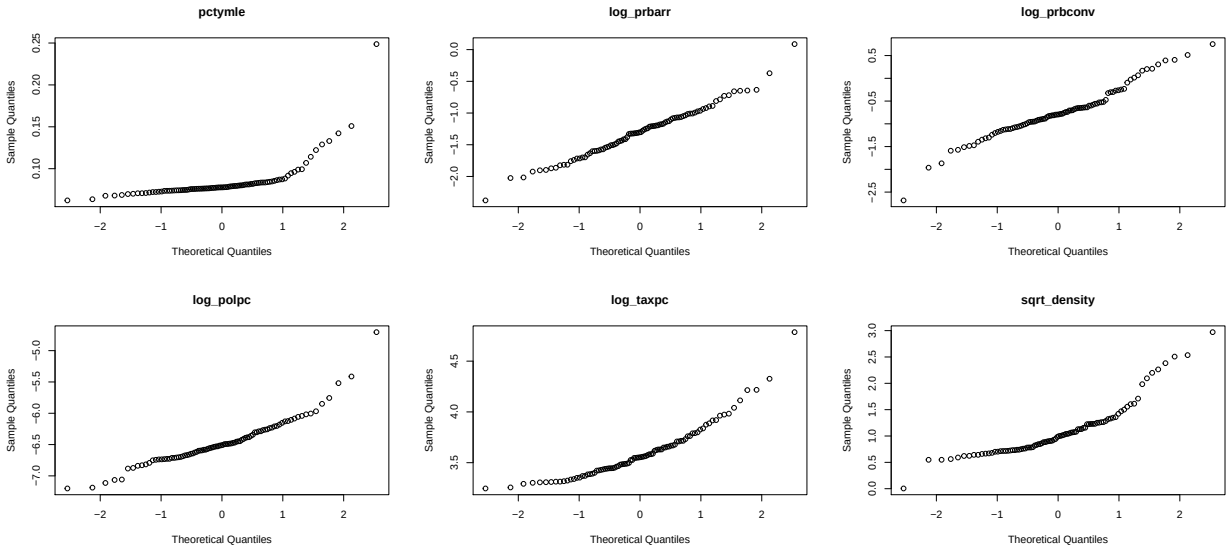
```
## # A tibble: 10 x 25
##    county  year  crmrte prbarr prbconv prbpris avgsen    polpc density
##     <int> <int>   <dbl>  <dbl>   <dbl>   <dbl>  <dbl>    <dbl>   <dbl>
##  1      3    87 0.0153   0.132    1.48   0.450   6.35 0.000746   1.05
##  2     19    87 0.0222   0.163    1.23   0.333  10.3  0.00202    0.577
##  3     99    87 0.0172   0.154    1.23   0.557  14.8  0.00186    0.548
##  4    115    87 0.00553  1.09     1.5    0.5    20.7  0.00905    0.386
##  5    127    87 0.0291   0.180    1.36   0.336  16.0  0.00158    1.34
##  6    137    87 0.0127   0.207    1.07   0.323   6.18 0.000814   0.317
##  7    149    87 0.0165   0.272    1.02   0.227  14.6  0.00152    0.609
##  8    185    87 0.0109   0.195    2.12   0.443   5.38 0.00122    0.389
##  9    195    87 0.0314   0.201    1.67   0.471  13.0  0.00446    1.75
## 10    197    87 0.0142   0.208    1.18   0.361  12.2  0.00119    0.890
## # ... with 16 more variables: taxpc <dbl>, west <int>, central <int>,
## #   urban <int>, pctmin80 <dbl>, wcon <dbl>, wtuc <dbl>, wtrd <dbl>,
## #   wfir <dbl>, wser <dbl>, wmfg <dbl>, wfed <dbl>, wsta <dbl>,
## #   wloc <dbl>, mix <dbl>, pctymle <dbl>
```

Detail: code to generate qqplots for evaluation of normality in EDA step

```
for (i in 1:length(colnames(crime_na))) {
    column_interest <- paste("crime_na$", colnames(crime_na)[i], sep = "")
    qqnorm(eval(parse(text = column_interest)), main = colnames(crime_na)[i])
}
```

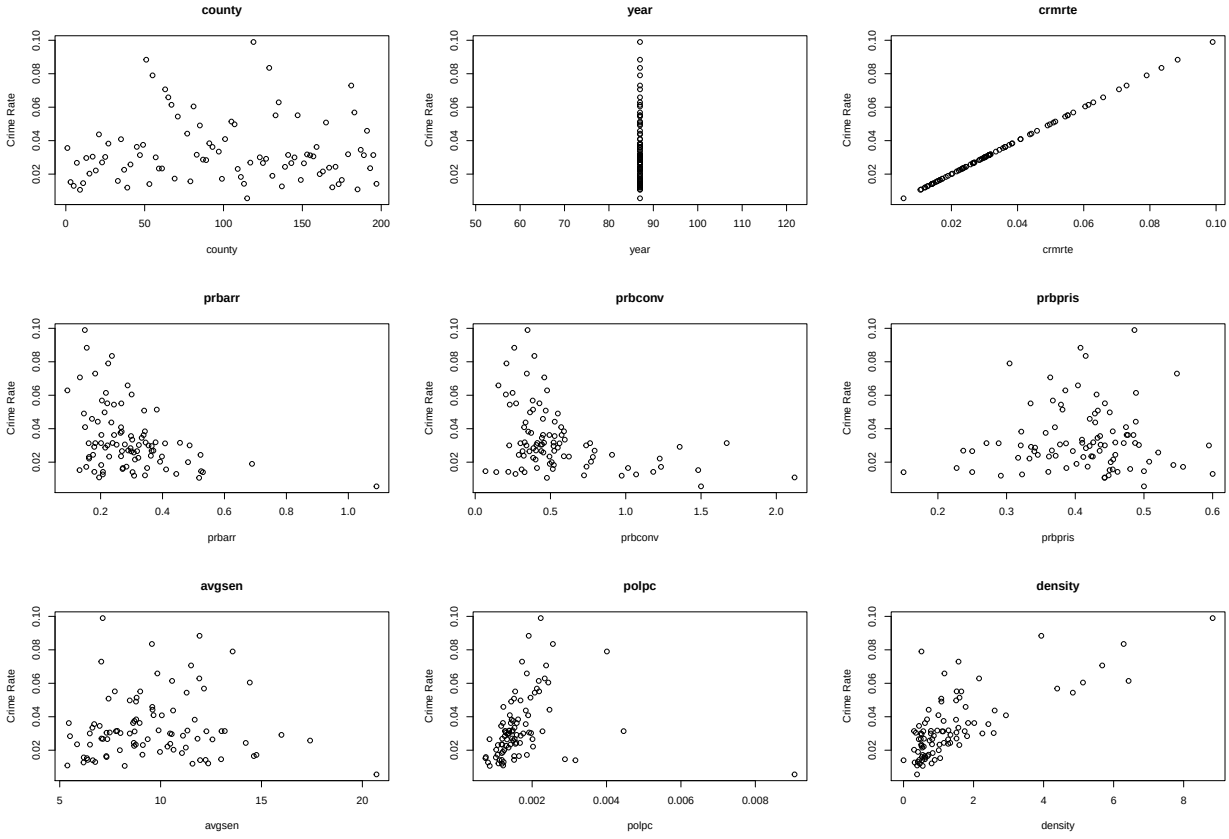


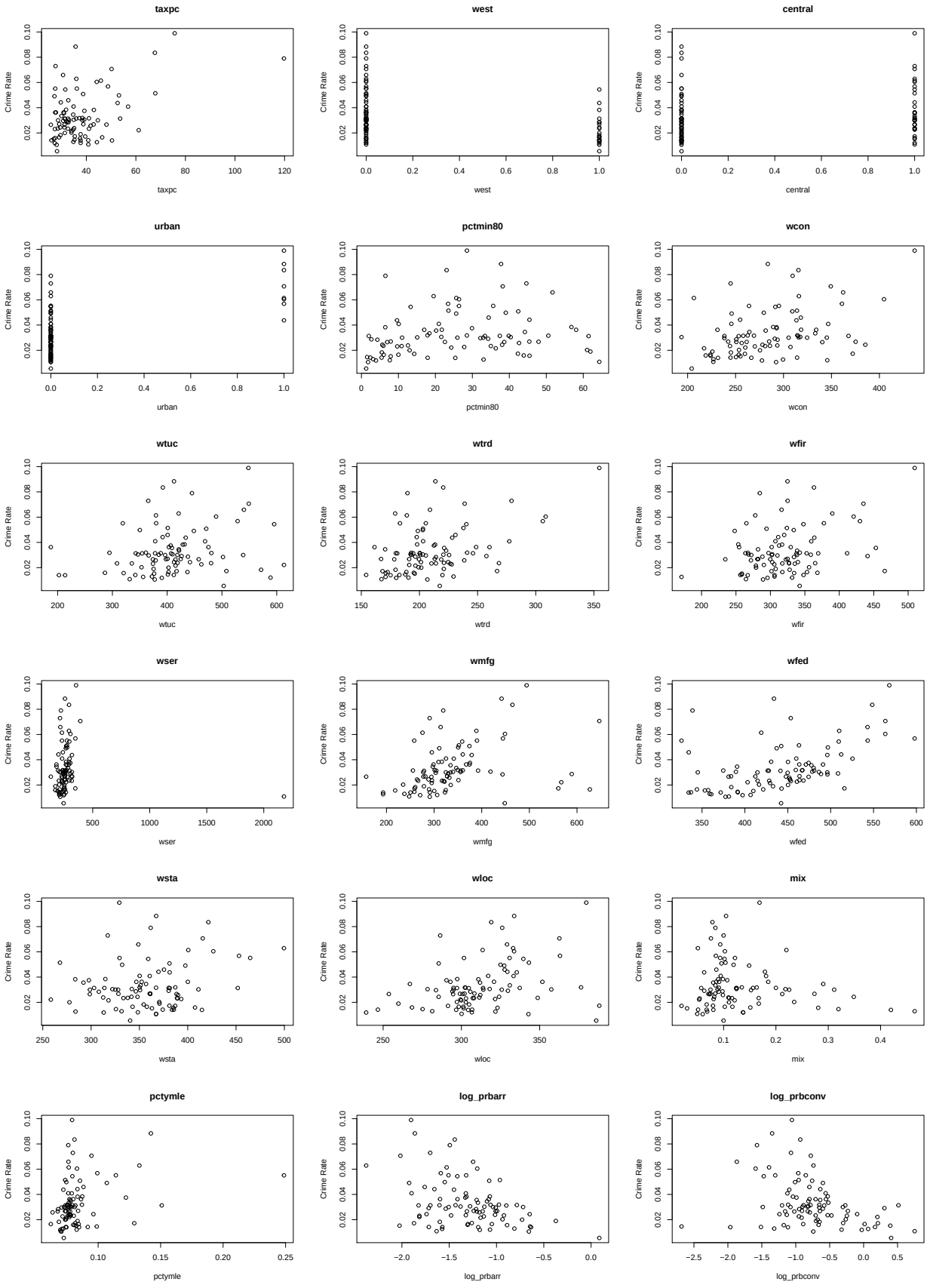

avgsen
polpc
density
taxpc
west
central
urban
pctmin80
wcon
wtuc
wtrd
wfir
wser
wmfg
wfed
wsta
wloc
mix
Sample Quantiles
Theoretical Quantiles

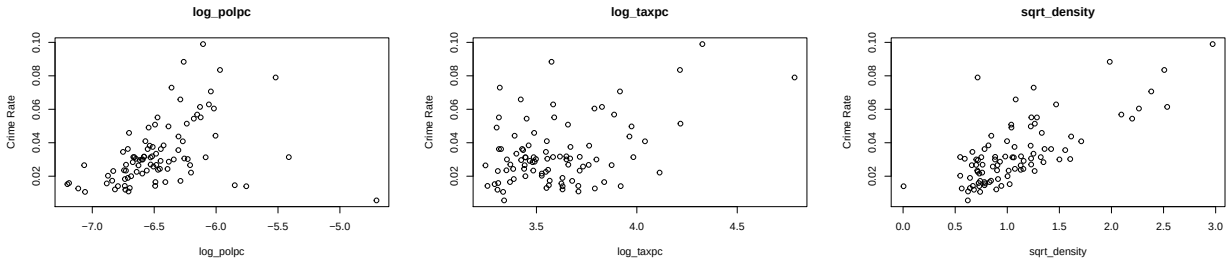


Detail: code to explore 'crime rate' vs. specific variables of interest

```
for (i in 1:length(colnames(crime_na))) {
    column_interest <- paste("crime_na$", colnames(crime_na)[i], sep = "")
    plot(eval(parse(text = column_interest)), crime_na$crmrte, main = colnames(crime_na)[i],
        ylab = "Crime Rate", xlab = colnames(crime_na)[i])
}
```

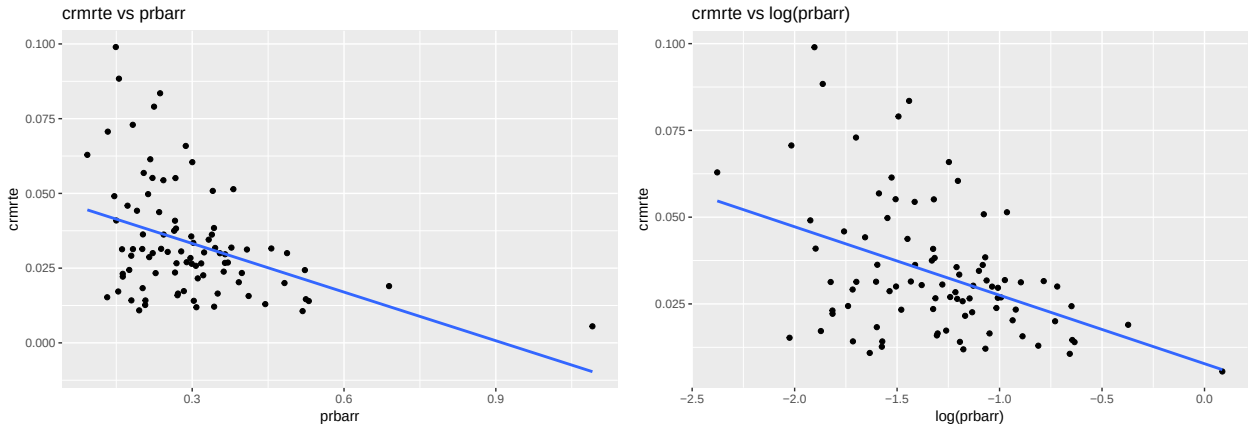



taxpc
west
central
urban
pctmin80
wcon
wtuc
wtrd
wfir
wser
wmfg
wfed
wsta
wloc
mix
pctymle
log_prbarr
log_prbconv
Crime Rate

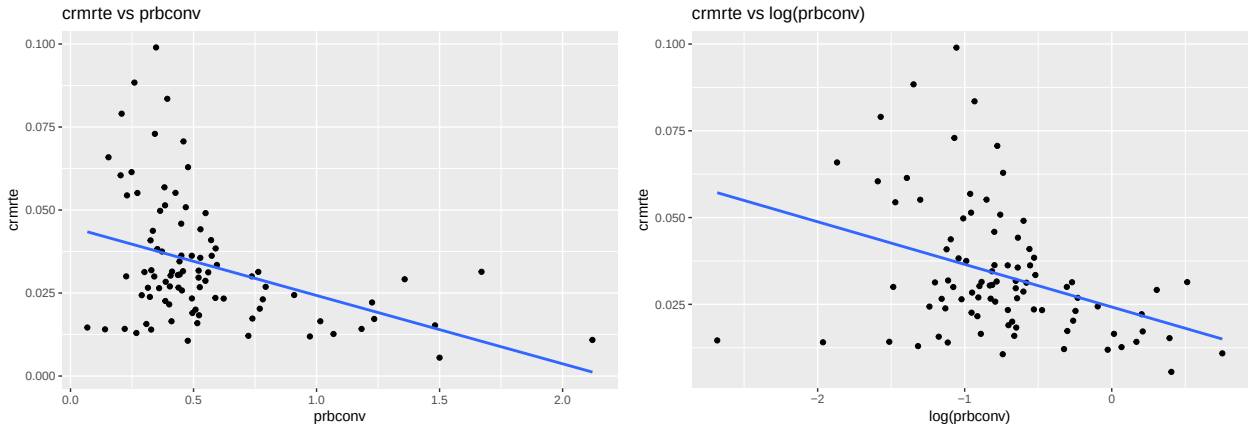


Detail: pre- and post-transformation scatterplots and q-q plots

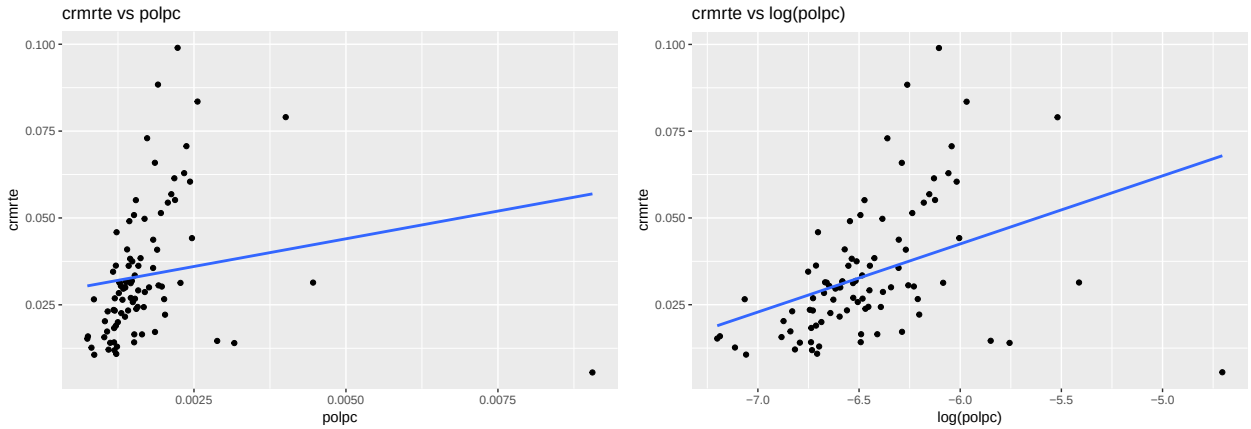
```
make_scatters(df = crime_na, var_list = c("prbarr"), y = "crmrte", trans = "log")
```



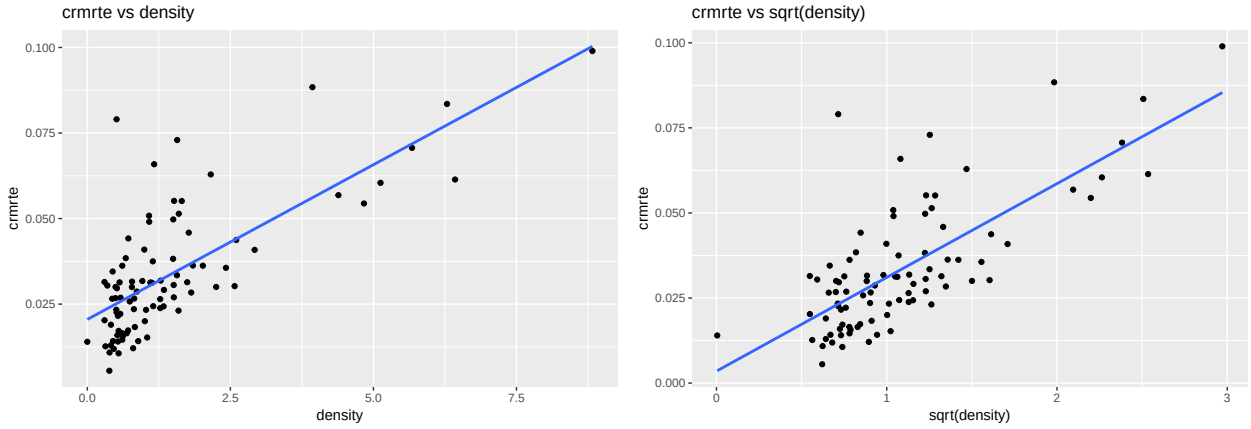

```
make_scatters(df = crime_na, var_list = c("prbconv"), y = "crmrte", trans = "log")
```



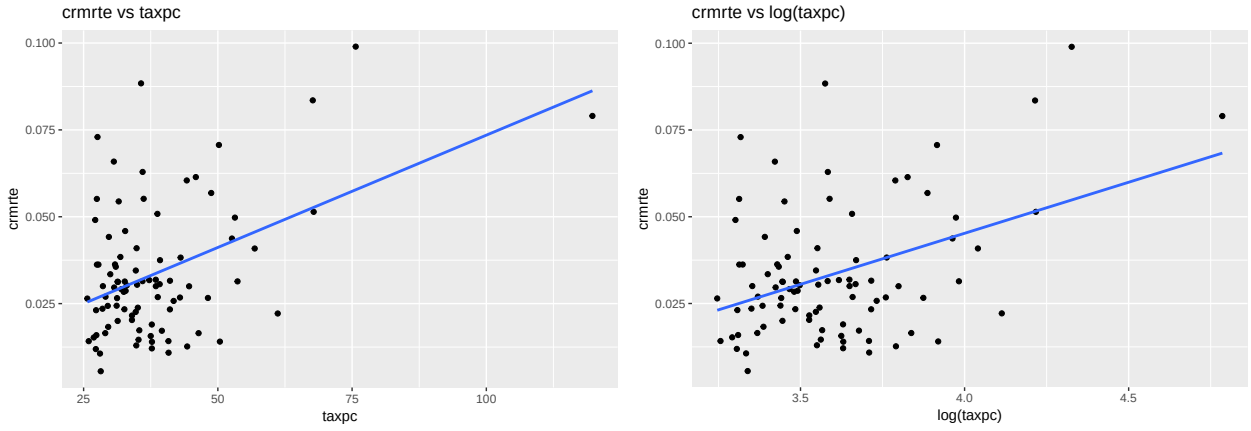

```
make_scatters(df = crime_na, var_list = c("polpc"), y = "crmrte", trans = "log")
```



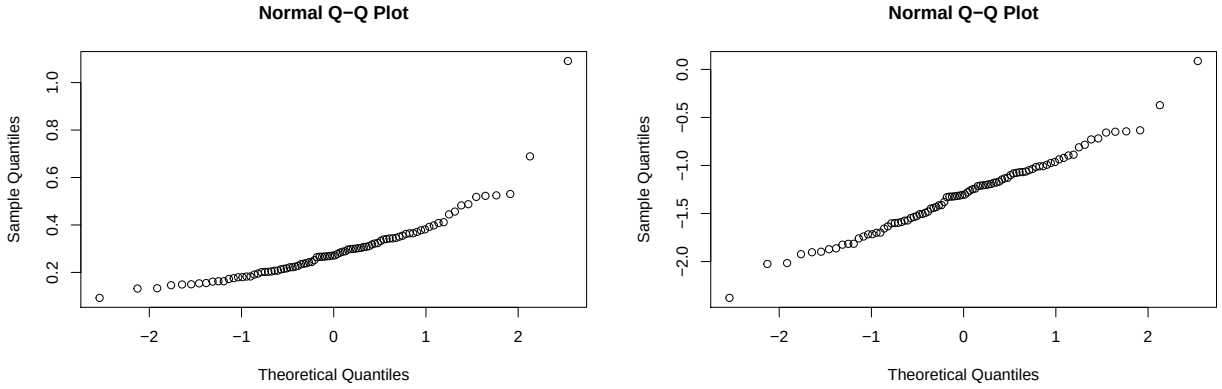

```
make_scatters(df = crime_na, var_list = c("density"), y = "crmrte", trans = "sqrt")
```



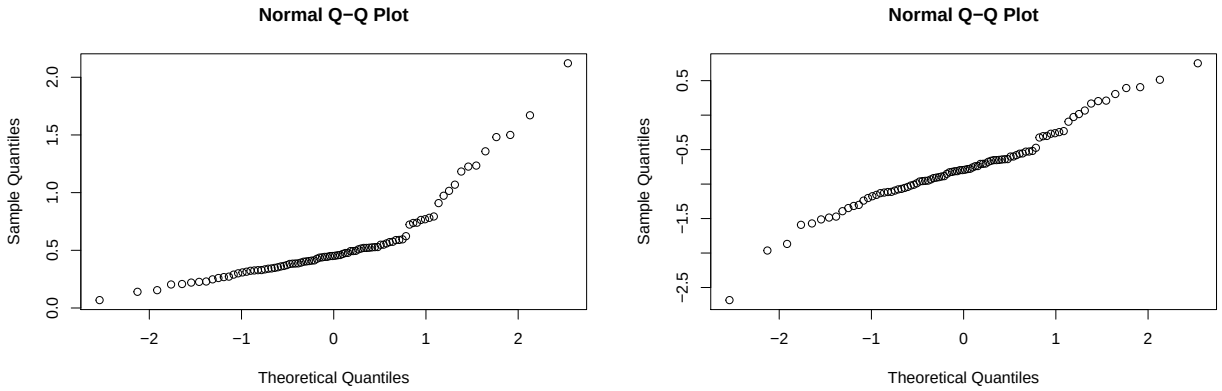

```
make_scatters(df = crime_na, var_list = c("taxpc"), y = "crmrte", trans = "log")
```



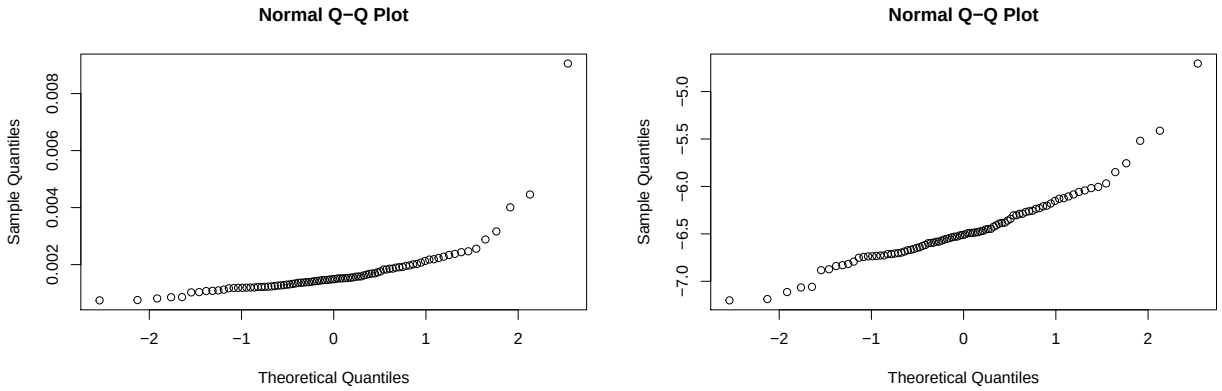

```
# Normality/skew with and without transformations
print(c(qqnorm(crime_na$prbarr), qqnorm(log(crime_na$prbarr))))
```



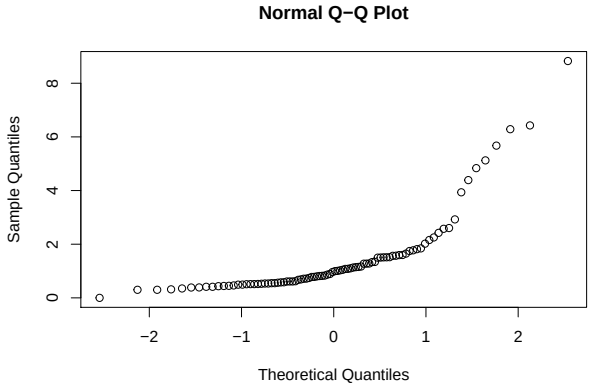

```
print(c(qqnorm(crime_na$prbconv), qqnorm(log(crime_na$prbconv))))
```

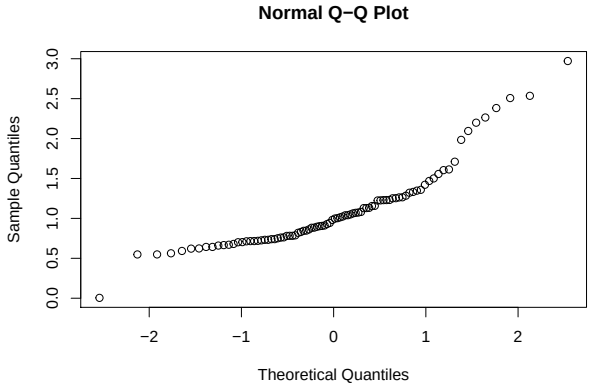



```
print(c(qqnorm(crime_na$polpc), qqnorm(log(crime_na$polpc))))
```




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






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