Crime in North Carolina

An Econometric Analysis of Data from 1987

Greg Tozzi

Contents

1.0 Introduction

The party's candidates in North Carolina need to address voters' concerns about crime with pragmatic, evidence-based policy prescriptions. The factors that likely have a causal relationship with the crime rate and that can be affected through policy are those related to the elements of the criminal justice system and those related to economic policy. Our research question is:

> Can we reduce the crime rate by putting policies in place to target income inequality or to adjust the efficiency or severity of the criminal justice system?

This study presents a model that seeks to explain North Carolina's crime rate based on county-level data from 1987. While the model is associative, it highlights certain variables that possibly have a causal relationship with the crime rate.

2.0 Loading and Cleaning the Data

The initial effort focuses on getting a sense of the data's structure and identifying and addressing anomalies.

2.1 Load Useful Packages and the Data

Load packages

```
suppressMessages(library(stargazer))
suppressMessages(library(tidyverse))
suppressMessages(library(car))
suppressMessages(library(lmtest))
suppressMessages(library(sandwich))
suppressMessages(library(kableExtra))
suppressMessages(library(psych))
suppressMessages(library(ggthemes))
suppressMessages(library(gridExtra))
suppressMessages(library(ggfortify))
source('nc_crime_helper_functions.R')
```

Load and review the data

Summary statistics build familiarity with the data and provide the first indications of anomalies.

```
crime <- read.csv('crime_v2.csv')
crime %>% summary() %>% formattedTable(5)
```

county	year	crmrte	prbarr	prbconv
Min. : 1	Min. :87	Min. :0.01	Min. :0.09	: 5
1st Qu.: 52	1st Qu.:87	1st Qu.:0.02	1st Qu.:0.21	0.588859022: 2
Median :105	Median :87	Median :0.03	Median :0.27	‘ : 1
Mean :102	Mean :87	Mean :0.03	Mean :0.29	0.068376102: 1
3rd Qu.:152	3rd Qu.:87	3rd Qu.:0.04	3rd Qu.:0.34	0.140350997: 1
Max. :197	Max. :87	Max. :0.10	Max. :1.09	0.154451996: 1
NA’s :6	NA’s :6	NA’s :6	NA’s :6	(Other) :86

prbpris	avgsen	polpc	density	taxpc
Min. :0.15	Min. : 5.38	Min. :0.00	Min. :0.00	Min. : 25.7
1st Qu.:0.36	1st Qu.: 7.34	1st Qu.:0.00	1st Qu.:0.55	1st Qu.: 30.7
Median :0.42	Median : 9.10	Median :0.00	Median :0.96	Median : 34.9
Mean :0.41	Mean : 9.65	Mean :0.00	Mean :1.43	Mean : 38.1
3rd Qu.:0.46	3rd Qu.:11.42	3rd Qu.:0.00	3rd Qu.:1.57	3rd Qu.: 40.9
Max. :0.60	Max. :20.70	Max. :0.01	Max. :8.83	Max. :119.8
NA’s :6	NA’s :6	NA’s :6	NA’s :6	NA’s :6

west	central	urban	pctmin80	wcon
Min. :0.00	Min. :0.00	Min. :0.00	Min. : 1.3	Min. :194
1st Qu.:0.00	1st Qu.:0.00	1st Qu.:0.00	1st Qu.: 9.8	1st Qu.:251
Median :0.00	Median :0.00	Median :0.00	Median :24.3	Median :281
Mean :0.25	Mean :0.37	Mean :0.09	Mean :25.5	Mean :285
3rd Qu.:0.50	3rd Qu.:1.00	3rd Qu.:0.00	3rd Qu.:38.1	3rd Qu.:315
Max. :1.00	Max. :1.00	Max. :1.00	Max. :64.3	Max. :437
NA’s :6	NA’s :6	NA’s :6	NA’s :6	NA’s :6

wtuc	wtrd	wfir	wser	wmfg
Min. :188	Min. :154	Min. :171	Min. : 133	Min. :157
1st Qu.:375	1st Qu.:191	1st Qu.:287	1st Qu.: 230	1st Qu.:289
Median :407	Median :203	Median :317	Median : 253	Median :320
Mean :412	Mean :212	Mean :322	Mean : 276	Mean :336
3rd Qu.:443	3rd Qu.:225	3rd Qu.:345	3rd Qu.: 281	3rd Qu.:360
Max. :613	Max. :355	Max. :509	Max. :2177	Max. :647
NA’s :6	NA’s :6	NA’s :6	NA’s :6	NA’s :6

wfed	wsta	wloc	mix	pctymle
Min. :326	Min. :258	Min. :239	Min. :0.02	Min. :0.06
1st Qu.:400	1st Qu.:329	1st Qu.:297	1st Qu.:0.08	1st Qu.:0.07
Median :450	Median :358	Median :308	Median :0.10	Median :0.08
Mean :443	Mean :358	Mean :313	Mean :0.13	Mean :0.08
3rd Qu.:478	3rd Qu.:383	3rd Qu.:329	3rd Qu.:0.15	3rd Qu.:0.08
Max. :598	Max. :500	Max. :388	Max. :0.47	Max. :0.25
NA's :6	NA's :6	NA's :6	NA's :6	NA's :6

2.1 Identify and Address Anomalies

Notably, there is not immediate evidence of top- or bottom-coding in the data. There are, however, a number of anomalies.

Misclassified variable

The variable `prbconv` was incorrectly read in as a `factor` vector, likely because of the errant accent mark revealed in the call to `summary` above. The variable should be `numeric`. Converting the variable first to a `character` vector and then to a `numeric` vector will result in non-numeric observations being converted to `NA` entries.

```
crime$prbconv <- crime$prbconv %>%
  as.character() %>%
  as.numeric()
```

```
## Warning in function_list[[k]](value): NAs introduced by coercion
```

The variable `prbconv` is now a `numeric` vector.

Missing data

Every variable has `6` `NA` entries. Examining the observations with `NA` entries reveals that all of the missing data lie in six rows of the data. Since these observations contain no data, removing them will not harm the analysis.

```
naIndex <- (crime %>%
              is.na() %>%
              rowSums()) > 0
crime %>%
  filter(naIndex == T) %>%
  kable(format = 'latex', booktabs = T) %>%
  kable_styling(latex_options = c('striped', 'scale_down'))
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc	density	taxpc	west	central	urban	pctmin80	wcon	wtuc	wtrd	wfir	wser	wmfg	wfed	wsta	wloc	mix	pctymle
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

```
crime  <- crime %>%
  drop_na()
```

Following removal, there are 91 remaining rows contain 0 NA entries.

Duplicate observations

Duplicate observations yield no additional information and increase the likelihood that ordinary least squares (OLS) regression will produce biased estimates. There is 1 duplicated row in the data set. This is easily verified by examining the rows in question.

```
duplicated_rows   <- crime %>%
  duplicated()
duplicated_county <- crime[which(duplicated_rows), 'county']
crime %>%
  filter(county == duplicated_county) %>%
  select(1:8) %>%
  kable(format = 'latex', booktabs = T) %>%
  kable_styling(latex_options = c('striped'))
```

county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc
193	87	0.024	0.266	0.589	0.423	5.86	0.001
193	87	0.024	0.266	0.589	0.423	5.86	0.001

Removing the duplicated row is trivial.

```
crime <- crime %>%
  distinct()
```

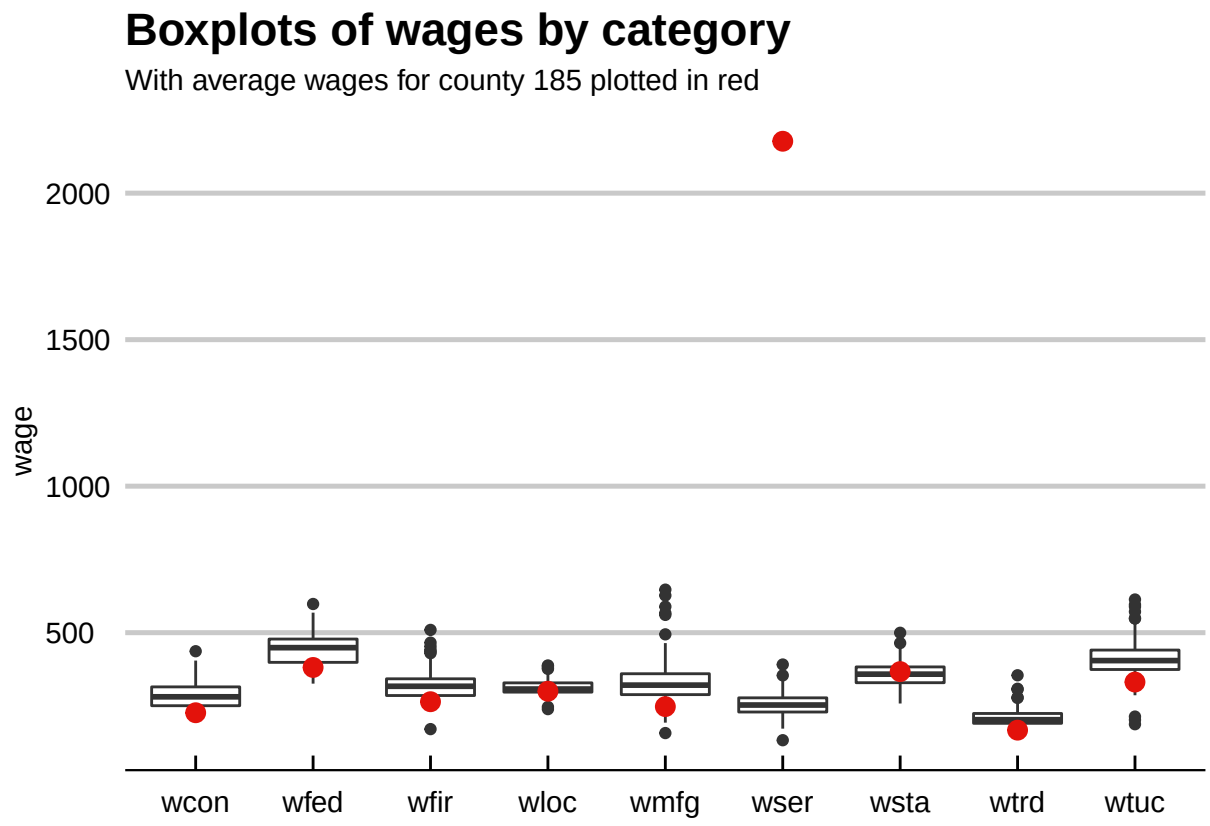
Following the removal, there are 90 observations and 0 duplicated rows.

Potential outlier in wser

The maximum value of wser is 2177, an order of magnitude larger than the median value of $253 and is, therefore, worthy of further examination. Constructing box plots of the various wage categories gives a sense of the distribution of average wages by category. Overlaying the plots with the average wages by category for the county with the maximum value of wser, county 185 makes clear that this is a unique outlier in the data. It is likely, although it cannot be demonstrated through the data alone, that the maximum value of wser was the result of a data entry error.

```
wageLongForm <- crime %>%
  select('county', wcon, wtuc, wtrd, wfir, wser, wmfg, wfed, wsta, wloc) %>%
  pivot_longer(-county, names_to = 'category', values_to = 'wage')
wserMaxIndex <- crime$wser %>%
  which.max()
wserCounty   <- crime$county[wserMaxIndex]
countyWages  <- wageLongForm %>%
  filter(county == wserCounty)
wagePlot     <- ggplot(wageLongForm, aes(x = category, y = wage)) +
  geom_boxplot() +
```

```
  theme_economist_white(gray_bg = FALSE) +
  geom_point(data = countyWages, colour = '#e3120b', size = 3) +
  xlab(NULL) +
  ggtitle(label = 'Boxplots of wages by category',
          subtitle = paste('With average wages for county', wserCounty, 'plotted in red'))
wagePlot
```

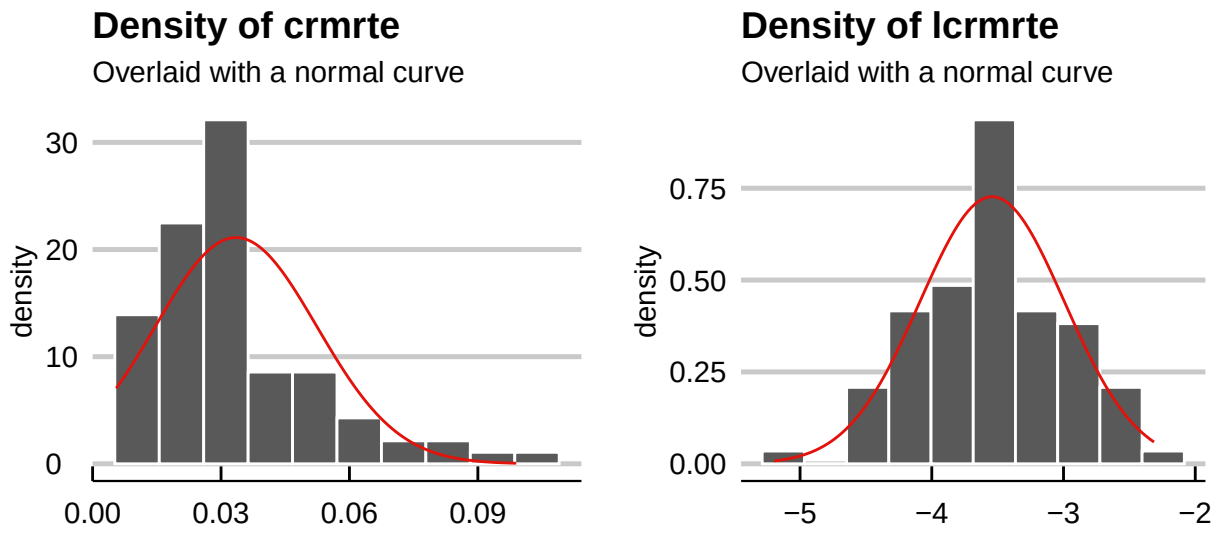


3.0 The Model Building Process

3.1 Univariate Analysis of the Outcome Variable

A challenge in building interpretable models around the crime rate is that the scale of the crime rate is likely not well understood outside of domain experts. It is more natural to think of the outcome of policy interventions in terms of the percent change in the crime rate. A logarithmic transformation of `crmrte` facilitates this. The logarithmic transformation is valid because the crime rate has a minimum value greater than zero and an unbounded maximum. Reviewing histograms of the untransformed and transformed variables reveals that the the positive skew in `crmrte` is reduced significantly with the transformation.

```
crime <- addLogColumns(crime, 3:ncol(crime))
crmrte_plot <- econHist('crmrte', crime)
lcrmrte_plot <- econHist('lcrmrte', crime)
grid.arrange(crmrte_plot, lcrmrte_plot, ncol = 2)
```



3.2 Implementation of the income inequality variable

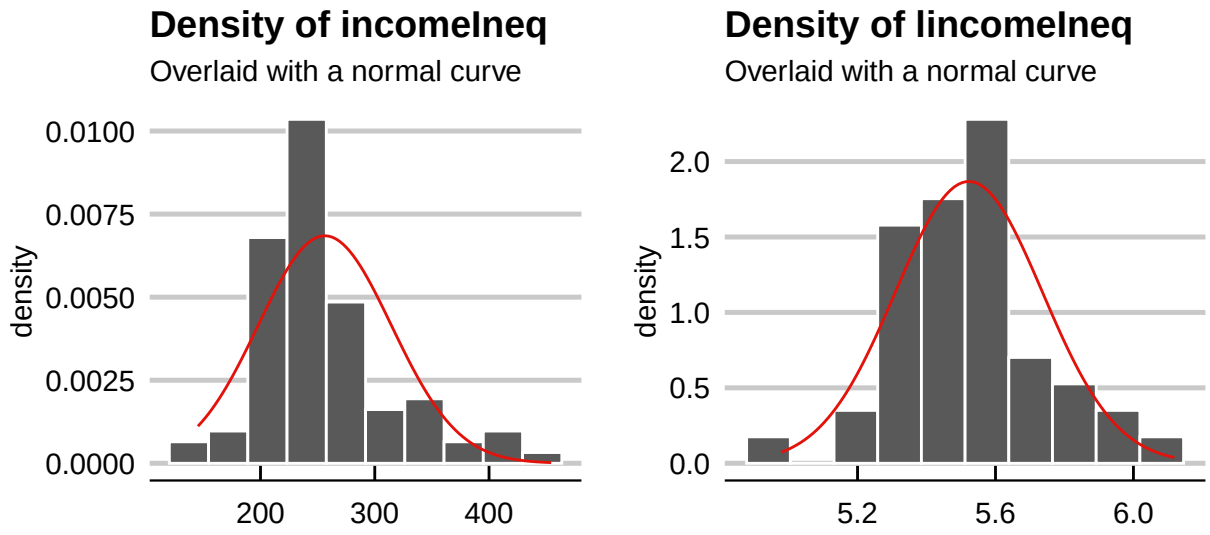
The research question requires a variable that serves as a proxy for income inequality. The data do not contain information about the spread of income between the most and least affluent residents of each county. The data do, however provide average wages by job category. A proxy, albeit imperfect, for income inequality is the range of the average wages by category. Unfortunately, as described above, there is likely a data entry error in the value of `wser` for county `185`. As a result, `wser` is not considered when implementing the income inequality variable.

```
wageColumns       <- c('wcon', 'wtuc', 'wtrd', 'wfir',
                       'wmfg', 'wfed', 'wsta', 'wloc')
maxWage           <- apply(crime[,wageColumns], 1, max) # Find the max average wage by county
minWage           <- apply(crime[,wageColumns], 1, min) # Find the min average wage by county
crime$incomeIneq  <- maxWage - minWage
crime$lincomeIneq <- log(crime$incomeIneq)
ineqDF            <- as.data.frame(crime[,'incomeIneq'])
names(ineqDF)     <- 'incomeIneq'
ineqDF %>%
  summary %>%
  kable(format = 'latex', booktabs = T) %>%
  kable_styling(latex_options = 'striped')
```

incomeIneq
Min. :145
1st Qu.:217
Median :248
Mean :256
3rd Qu.:273
Max. :454

The table above suggests that the `incomeIneq` is skewed. The histogram of `incomeIneq` confirms this suspicion. A logarithmic transformation renders the distribution of `incomeIneq` somewhat closer to normal.

```
incomeIneq_plot  <- econHist('incomeIneq', crime)
lincomeIneq_plot <- econHist('lincomeIneq', crime)
grid.arrange(incomeIneq_plot, lincomeIneq_plot, ncol = 2)
```



3.3 Base Model

The first model only includes the key explanatory variables that direct address the research question. Those variables that are considered :

1. **The crime rate**. The transformed variable, `lcrmrte` is preferred to the untransformed `crmrte` for the reasons described in section 3.1.

2. **Income inequality**. This is proxied as described in 3.2. Income inequality can be addressed through policy by raising the minimum wage and supporting collective bargaining by labor unions. We hypothesize that increased income inequality will be associated with an increase in the crime rate.

3. **Probability of arrest**. The probability of arrest serves as an indicator of the aggressiveness of law enforcement. The variable is proxied as the ratio of the number of arrests per reported crime. Values greater than 1.0 are plausible in that arrests can be effected by a jurisdiction for crimes that occurred outside of that jurisdiction and arrests can be made in the current calendar year for crimes reported in previous calendar years. The probably of arrest can be affected by adjusting police resourcing, training, and tactics. We hypothesize that an increase in the probability of arrest will be associated with a decrease in the crime rate.

4. **Probability of conviction**. This is a measure of the effectiveness of investigative and prosecutorial efforts proxied by the ratio of convictions per arrest. Values greater than 1.0 are plausible in that a single criminal may face several charges at a trial following a single arrest and convictions may occur in the current calendar year for arrests that occurred in the previous calendar year. The probability of conviction can be affected by adjusting training and resources for investigators and prosecutors or by redefining crimes through statutory changes. We hypothesize that an increase in the probability of conviction will be associated with a decrease in the crime rate.

5. **Probability of prison**. The likelihood that an individual will be sentenced to prison, proxied by the ratio of convictions resulting in a prison sentence to total convictions, gives a sense of the willingness of society to impose punishments involving the loss of liberty. The probability of a prison sentence can

be affected through changes to the criminal statute and sentencing guidelines. We hypothesize that an increase in the probability of being sentenced to prison will be associated with a reduction in the crime rate.

6. **Average prison sentence**. The average prison sentence is a measure of the severity of punishment. The data provide the average prison sentence in days. We hypothesize that an increase in the average sentence will be associated with a reduction in the crime rate.

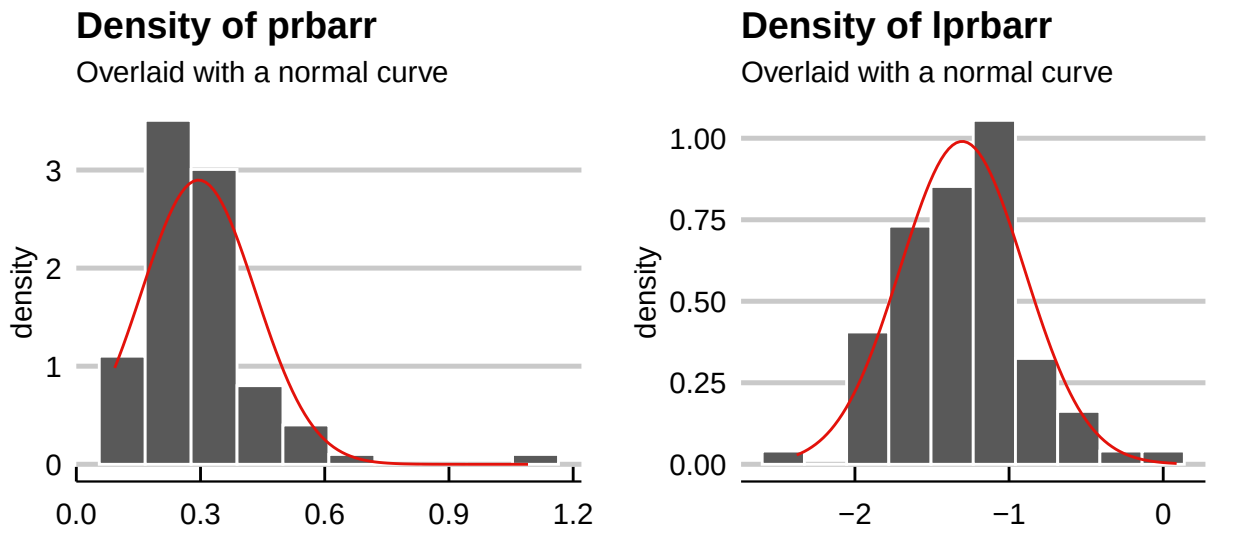
7. **Per-capita number of police**. This is a measure of the resources available to law enforcement. We hypothesize that an increase in the number of police per capita will be associated with a reduction in the crime rate.

Univariate analysis of explanatory variables

The distributions of `crmrte`, `incomeIneq`, and their logarithmic transformations were explored in the histograms in sections 3.1 and 3.2. The histograms that follow that show the distributions of `prbarr`, `probconv`, `prbpris`, `avgsen`, `polpc`, and their logarithmic transformations.

The histogram of `prbpris` shows a strong positive skew. This is substantially corrected with the logarthimc transformation. The maximum value of `prbarr` appears to be an outlier. The value, `1.091`, is associated with county `115`. We will continue to collect interesting outliers to be examined at the end of the univariate analysis.

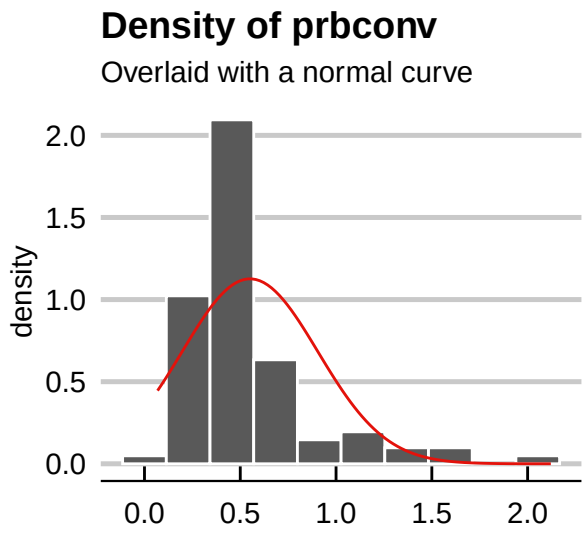
```
baseAndLogPHist('prbarr', crime)
```

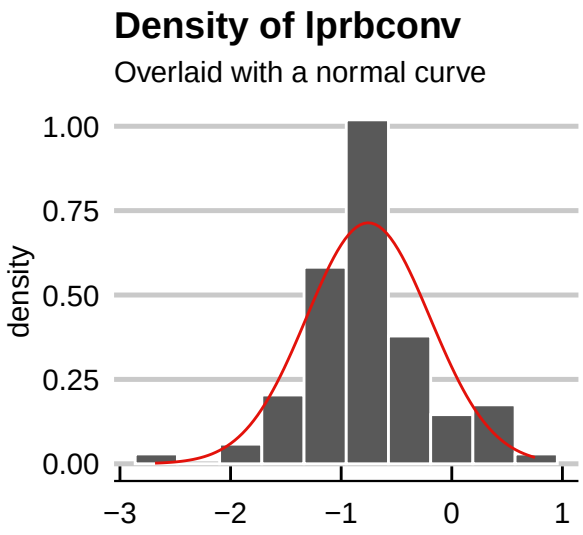



```
interestingCounties <- c(crime$county[which.max(crime$prbarr)])
```

Likewise, the histogram of `prbconv` shows a positive skew that is corrected with the logarithmic transformation. The maximum and minimum values of `prbconv` are interesting outliers. The minimum value is `0.068` and is associted with county `11`. The maximum value is `2.121` associated with county `185`.

```
baseAndLogPHist('prbconv', crime)
```

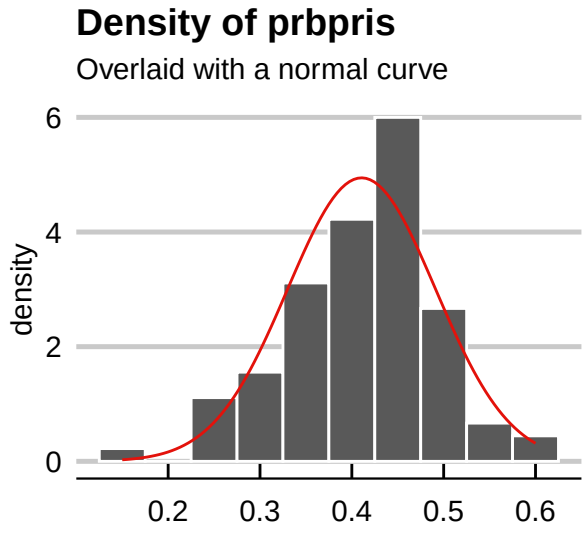


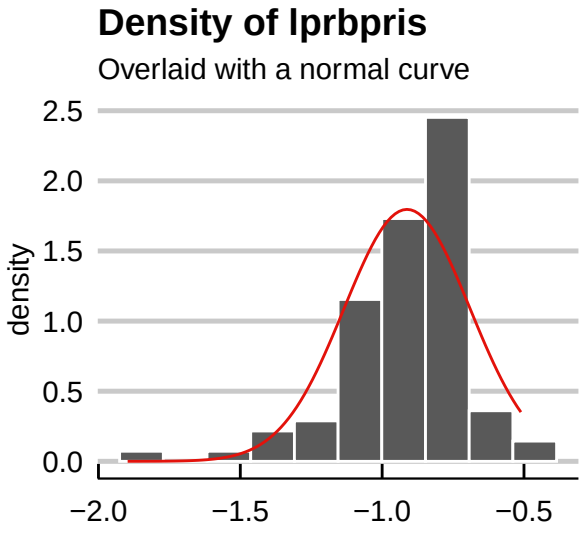



```
interestingCounties <- union(interestingCounties,
                             c(crime$county[which.min(crime$prbconv)],
                               crime$county[which.max(crime$prbconv)]))
```

The histogram of `prbpris` shows negative skew. As expected, the logarithmic transformation exacerbates this skew.

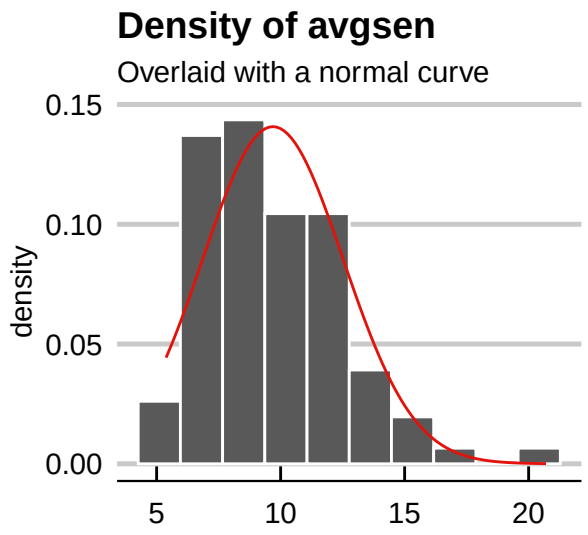
```
baseAndLogPHist('prbpris', crime)
```

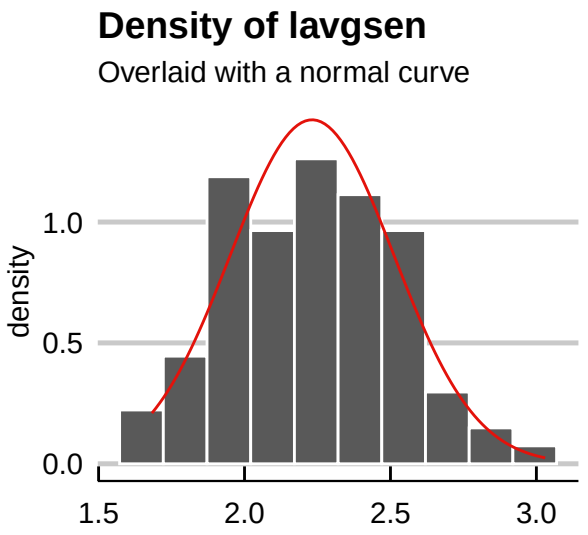




The average sentence is positively skewed. The maximum value is potentially interesting as an outlier. Its value is `20.7` associated with county `115`.

```
baseAndLogPHist('avgsen', crime)
```

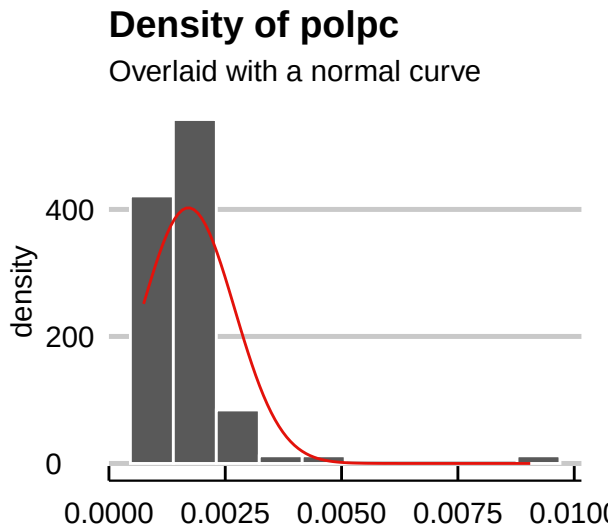


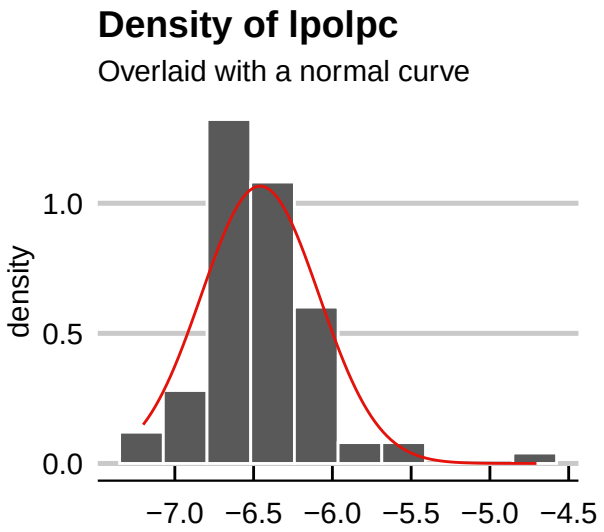



```
interestingCounties <- union(interestingCounties,
                             crime$county[which.max(crime$avgsen)])
```

The number of police per capita has a large positive skew driven in large part by a single large excursion from the mean. This value of `0.009` is associated with county `115`.

```
baseAndLogPHist('polpc', crime)
```






```
interestingCounties <- union(interestingCounties,
                             crime$county[which.max(crime$polpc)])
```

The potential outliers listed above are worthy of a quick review. A series of box plots of the standardized variables gives a sense of the counties that produced these values. Start by standardizing the variables of interest.

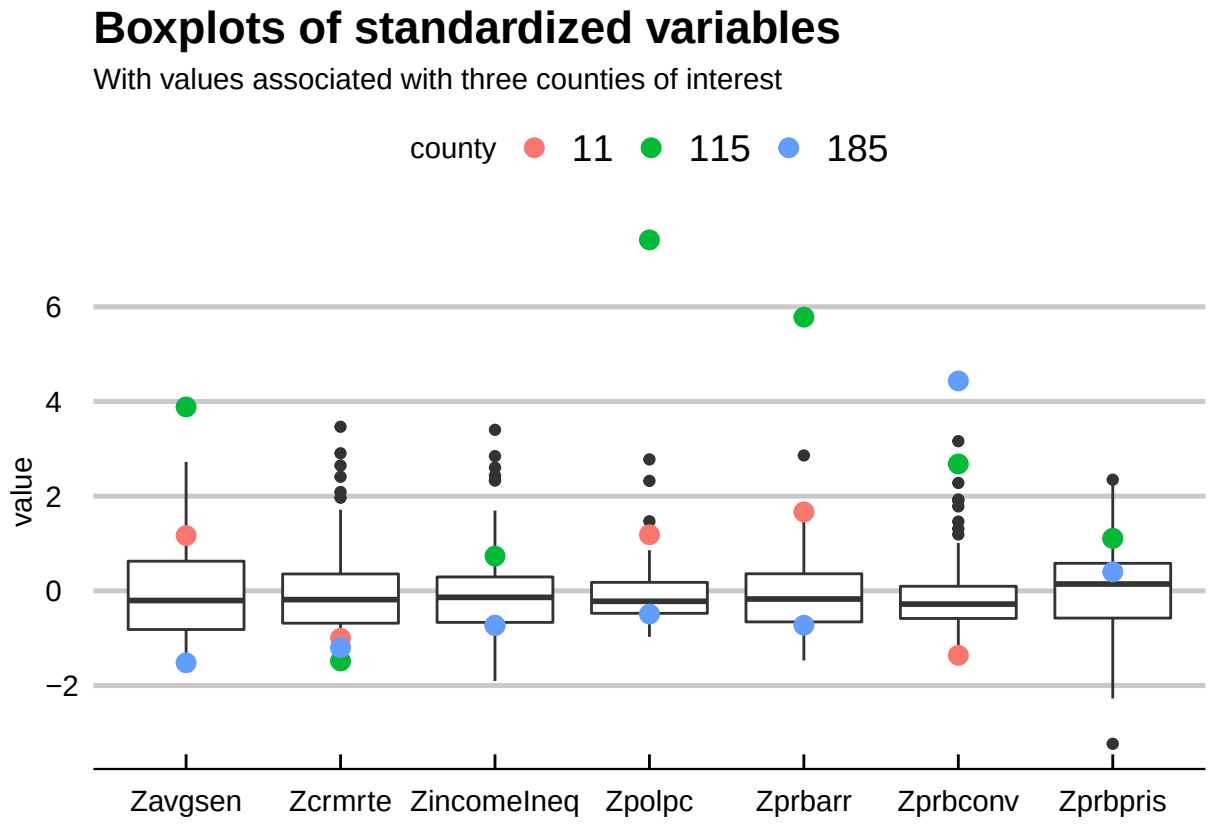
```
crimeZLong <- crime %>%
  mutate(
    Zcrmrte     = Z(crmrte),
    Zprbarr     = Z(prbarr),
    Zprbconv    = Z(prbconv),
    Zprbpris    = Z(prbpris),
    Zavgsen     = Z(avgsen),
    Zpolpc      = Z(polpc),
    ZincomeIneq = Z(incomeIneq)) %>%
  select(county, Zcrmrte, Zprbarr,
         Zprbconv, Zprbpris, Zavgsen,
         Zpolpc, ZincomeIneq) %>%
  pivot_longer(-county,
               names_to = 'category',
               values_to = 'value')
crimeZLong$county <- as.factor(crimeZLong$county)
```

Then find the values of each variable for the counties of interest.

```
interestingVals <- crimeZLong %>%
  filter(county %in% interestingCounties)
```

Now construct the box plots.

```
outlierPlot <- ggplot(crimeZLong,
                      aes(x = category, y = value)) +
  geom_boxplot() +
  theme_economist_white(gray_bg = FALSE) +
  geom_point(data = interestingVals,
             aes(colour = county),
             size = 3) +
  xlab(NULL) +
  scale_fill_manual(name = 'County') +
  ggtitle(label = 'Boxplots of standardized variables',
          subtitle = 'With values associated with three counties of interest')
outlierPlot
```

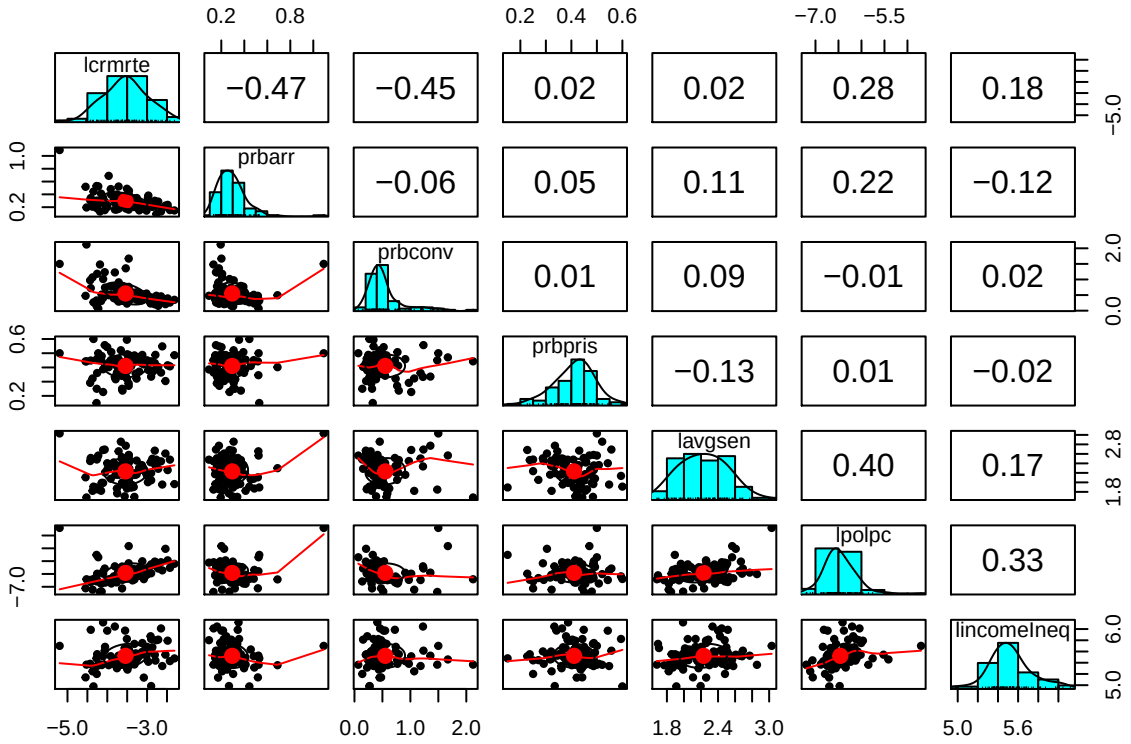


The box plot above reveals three counties with different approaches to law enforcement that all lead to relatively low crime rates. County 11's relatively large law enforcement contingent is effective at making arrests. It's investigators and prosecutors, however, are relatively ineffective at bringing cases to successful outcomes. County 115's relatively massive police force makes a relatively large number of arrests that typically result in convictions and long prison sentences. County 185's small police force produces unremarkable arrest statistics but brings cases forward that result in convictions. These convictions, however, result in short sentences. While these three narratives are interesting, nothing in them seems to suggest an *a priori* reason to discount the values associated with counties 11, 115, and 185.

Multivariate analysis

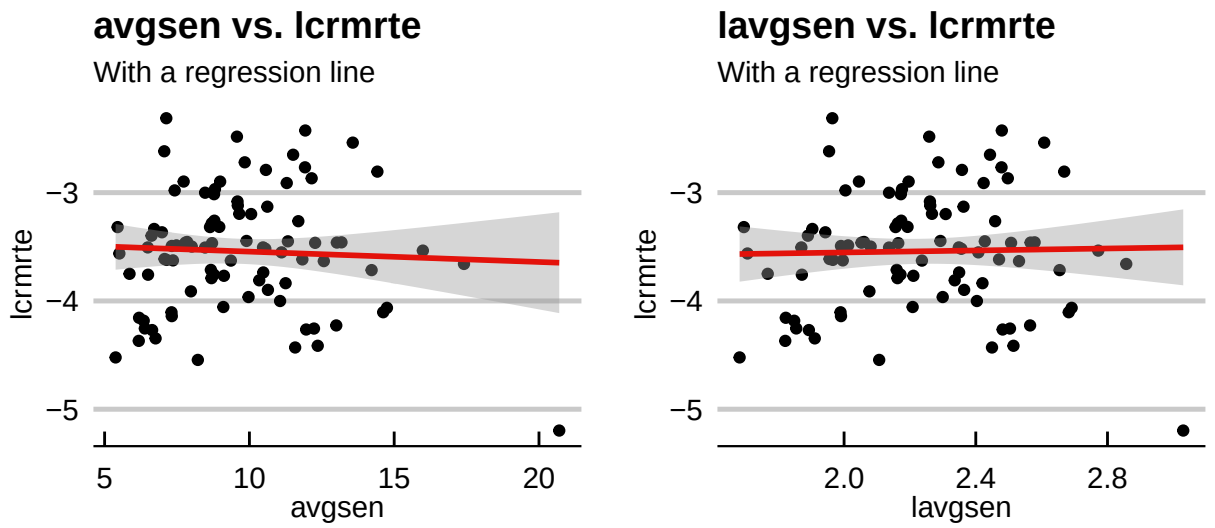
A scatterplot matrix reveals relationships between the outcome and the explanatory variables explored in the univariate analysis above. Logarithmic transformations were considered for all variables that showed a positive skew. Where a logarithmic transformation would make interpretability challenging, as in the cases of those variables that are proxies for percentages, the transformation was not applied.

```
crime_1 <- crime %>% select(lcrmrte, prbarr, prbconv,
                            prbpris, lavgsen, lpolpc,
                            lincomeIneq)
pairs.panels(crime_1)
```



Correlations between `lcrmrte` and the probabilities of arrest and conviction are negative, consistent with the hypotheses for the effect of these variables. Interestingly, the probability of prison and the transformed average sentence seem to have little impact on the crime rate. The scatterplots below confirm that the lack of correlation is not an artifact of transforming `avgsen`.

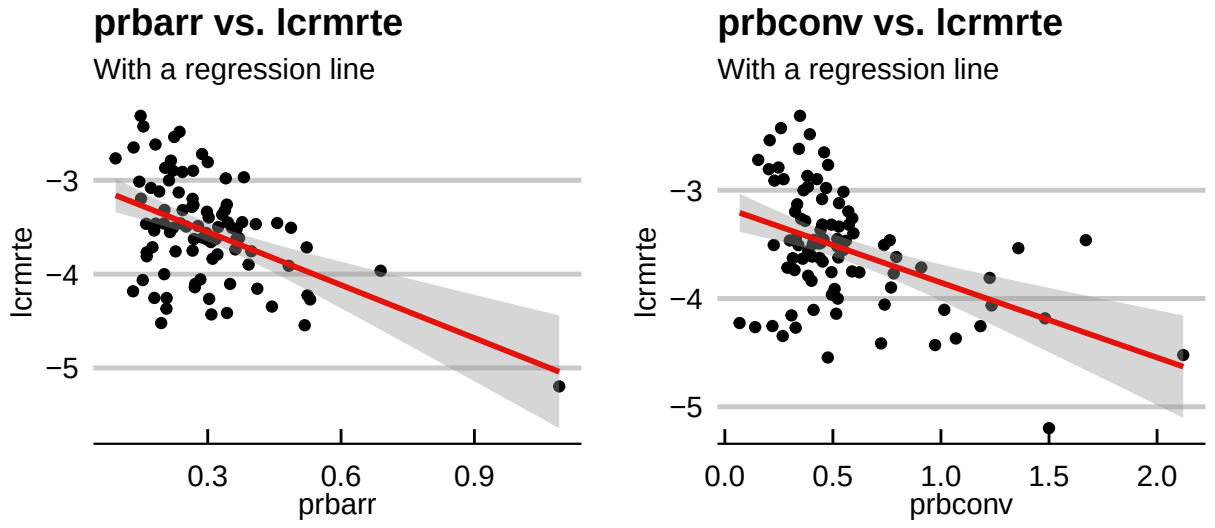
```
avgsen_scatter  <- econPlot('avgsen', 'lcrmrte', crime)
lavgsen_scatter <- econPlot('lavgsen', 'lcrmrte', crime)
grid.arrange(avgsen_scatter, lavgsen_scatter, ncol = 2)
```



The positive correlation between `lcrmrte` and the transformed measure of income inequality is consistent with the hypothesis that an increase in income inequality is associated with an increase in the crime rate. The positively correlation between `polpc` and `lcrmrte`, however, is counterintuitive. It suggests that an increase in the relative size of the police force is associated with an increase in the crime rate. It may be the case that the causal flow is such that increases in the crime rate are met quickly by increasing the end strength of the police force.

The scatterplot matrix above does not offer an adequate resolution to determine the nature of the relationships between the independent variables and `lcrmrte`. These relationships are explored in the scatterplots that follow.

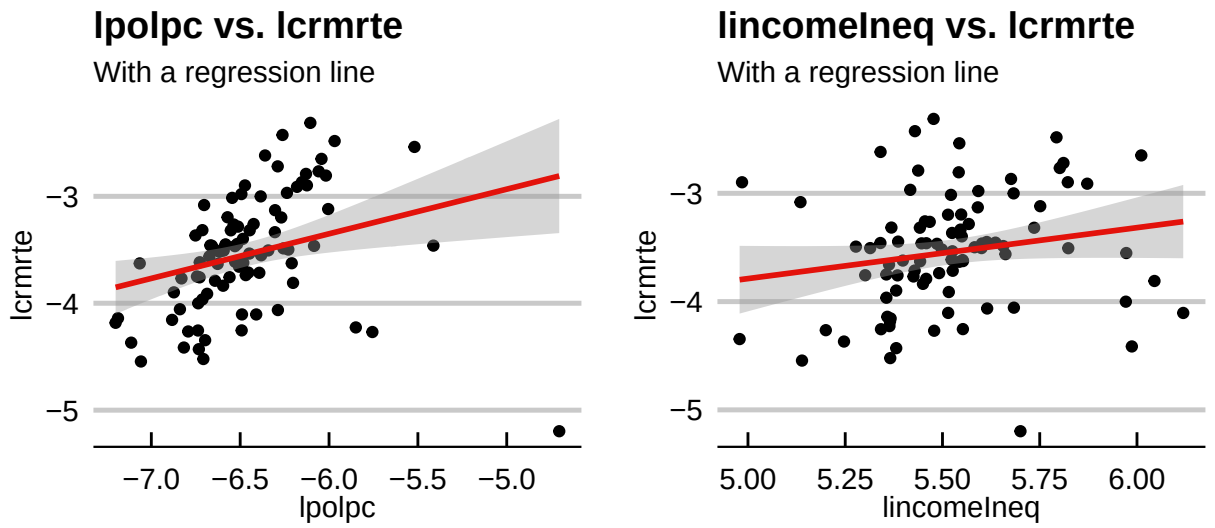
```
prbarr_scatter  <- econPlot('prbarr', 'lcrmrte', crime)
prbconv_scatter <- econPlot('prbconv', 'lcrmrte', crime)
grid.arrange(prbarr_scatter, prbconv_scatter, ncol = 2)
```



The plots above suggest that the moderate negative correlations with `lcrmrte` for the variables `prbarr` and `prbconv` may be the result of the outlier values demonstrated above to be associated with counties 115 and

185. Care will be taken to examine the effects of outliers in the regressions that will follow.

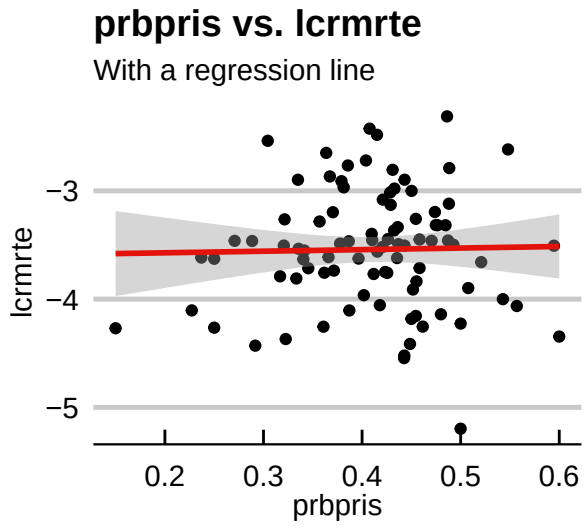
```
lpolpc_scatter      <- econPlot('lpolpc', 'lcrmrte', crime)
lincomeIneq_scatter <- econPlot('lincomeIneq', 'lcrmrte', crime)
grid.arrange(lpolpc_scatter, lincomeIneq_scatter, ncol = 2)
```



The plots above both show some evidence of nonlinearity. In particular, the relationship between `lcrmrte` and `lpolpc` may be quadratic. Unfortunately, further transformations of `polpc` are likely to render its effect on the crime rate extremely challenging to interpret.

For completeness, the scatterplot below examines the relationship between `lavgsen` and `prbpris`. Contrary to the hypothesis expressed in 3.3, there is no obvious relationship.

```
prbpris_scatter <- econPlot('prbpris', 'lcrmrte', crime)
grid.arrange(prbpris_scatter, ncol = 2)
```



Initial model specification

The initial specification is:

$$log(crmrte) = \beta_0 + \beta_1 prbarr + \beta_2 prbconv + \beta_3 prbpris + \beta_4 log(avgsen) + \beta_5 log(polpc) + \beta_6 log(incomeIneq) + u$$

Fitting the model

We fit the model using OLS and examine the coefficients.

```
fit_1 <- lm(lcrmrte ~ prbarr + prbconv + prbpris +
              lavgsen + lpolpc + lincomeIneq, data = crime)
fit_1r2 <- summary(fit_1)$r.squared
fit_1CoeffDF <- fit_1$coefficients %>% as.data.frame()
names(fit_1CoeffDF) <- 'coefficients'
fit_1CoeffDF %>% kable(format = 'latex', booktabs = T) %>%
  kable_styling()
```

	coefficients
(Intercept)	1.846
prbarr	-2.360
prbconv	-0.734
prbpris	0.306
lavgsen	-0.063
lpolpc	0.626
lincomeIneq	-0.042

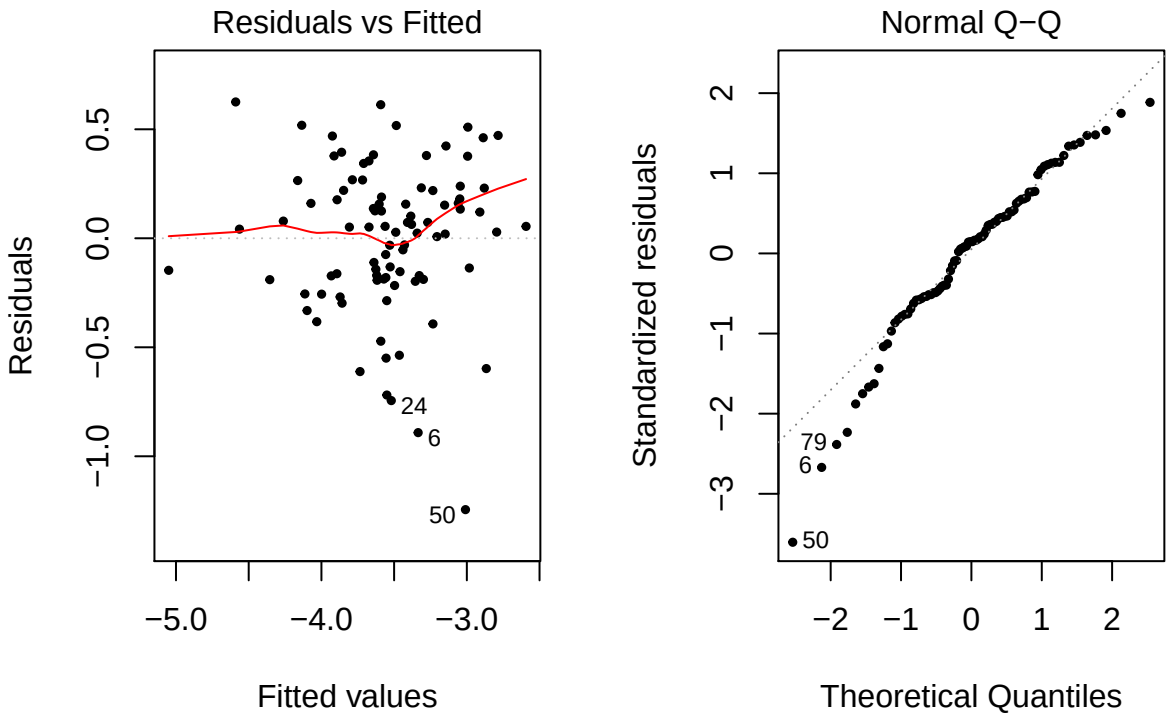
The base model's R^2 is `0.61` meaning that the model's independent variables explain `61`% of the variance.

Highlights of evaluation of the OLS assumptions

It is not clear that the model meets the assumption of random sampling. While temporal autocorrelation is clearly not a factor, *spatial autocorrelation* needs to be assessed. It stands to reason that the observations for adjacent counties may be correlated. We do not have the tools to assess spatial autocorrelation and will leave this matter for further study.

The residuals vs. fitted values plot below gives some evidence of a violation of the zero conditional mean assumption. It appears that the excursions from zero, while small, increase toward the right side of the plot. We will accept this for the time being but will seek to address it in subsequent models.

```
par(mfrow = c(1, 2))
plot(fit_1, which = c(1,2), pch = 19, cex = 0.5)
```



A review of the Q-Q plot gives clear evidence that the the normality assumption has been violated. A Shapiro-Wilks test (p = `0.008`) provides further evidence that the residuals are not normally distributed. However, the sample size is sufficient that we can rely on OLS asymptotics and proceed with the analysis.

Assessing statistical significance

While a Breusch-Pagan test (p = `0.374`) suggests that we cannot reject the null hypothesis of homoscedasticity, we follow the best practice of computing heteroscedasticity-robust standard errors.

```
fit_1Coeftest <- coeftest(fit_1, vcov = vcovHC)
fit_1Coeftest %>% coefTestTable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.846	2.184	0.845	0.400
prbarr	-2.360	0.369	-6.398	0.000
prbconv	-0.734	0.102	-7.167	0.000
prbpris	0.306	0.739	0.414	0.680
lavgsen	-0.063	0.191	-0.331	0.741
lpolpc	0.626	0.148	4.220	0.000
lincomeIneq	-0.042	0.240	-0.174	0.862

Interpreting the coefficients

β_1. A unit change in the probability of probability of arrest is associated with a `-235.982`% change in the crime rate. A unit change in the probability of arrest is a 100% percent change, which is too large to be useful. It is perhaps more natural to state that increasing the probability of arrest by 0.01 (1%) is associated with a `-2.36`% change in the crime rate.

β_2. Following the form of the analysis above, increasing the probability of conviction by 0.01 (1%) is associated with a `-0.734`% change in the crime rate.

β_5. A 1% increase in the number of police per capita is associated with a `0.626`% increase in the crime rate.

β_3, β_4, and β_6 are not significantly different than zero, suggesting little association between the probability of receiving a prison sentence, changes in the average sentence length, or changes in our measure of income inequality with changes in the crime rate.

3.2 A Second Model

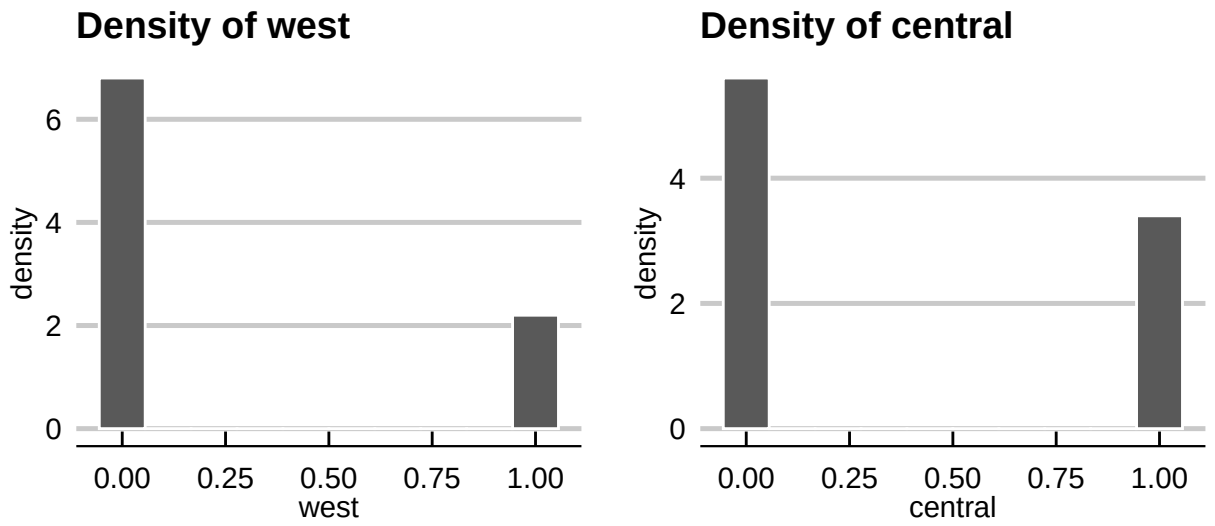
Adding variables may increase the accuracy of the model. The base model does not control for the following potentially interesting effects.

1. Demographic effects. The variable `pctymle` gives the percentage of young men aged 15-24 in a given county. Members of this demographic may be more likely to commit crime than other members of the population.
2. Effects related to urbanization. The nature of crime in urban areas is likely different form that in rural areas. A proxy for urbanization is the variable `density`. This variable provides the population density for each of the counties included in the data.
3. Regional effects. The nature of crime likely has something to do with where in the state a county lies. Beyond issues of urbanization, transportation connections may facilitate illicit activity in certain parts of the state. North Carolina's major axis runs east-west. The data split the state into a western, central, and–presumably–an eastern region. The west region is captured in the dummy variable `west`, the central in `central`.

Univariate analysis of the added explanatory variables

As with the base model, we turn our attention to the distributions of the new explanatory variables in order to determine which of them may benefit from transformation. The regional variables are binary in nature.

```
west_plot <- econHist('west',
                      crime,
                      curve = F)
central_plot <- econHist('central',
                         crime,
                         curve = F)
grid.arrange(west_plot,
             central_plot,
             ncol = 2)
```



The distributions above do not provide any indication of counties that are anomalously coded both `central` and `west`. Searching for such values is a sensible precaution.

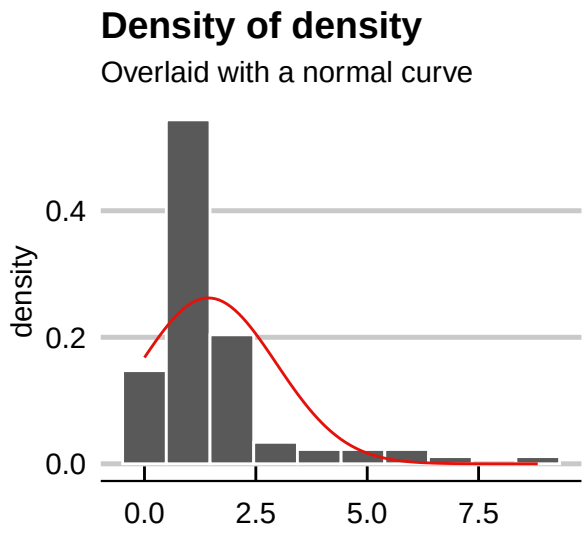
```
dualCounties <- crime %>%
  filter(west == 1 & central == 1) %>%
  select(county, west, central)
dualCounties %>%
  kable(format = 'latex', booktabs = T) %>%
  kable_styling()
```

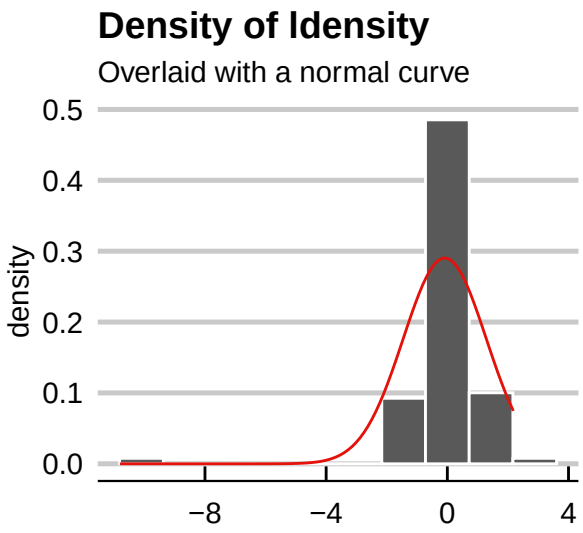
county	west	central
71	1	1

County `71` is coded as both `west` and `central`. This is likely a data entry error. Retaining the county has the benefit of not losing additional information related to the observation.

Plotted below are the histograms of `density` and its logarithmic transformation, `ldensity`. The variable's scale is almost certainly off by two orders of magnitude. The code book states that `density` reflects the number of people per square mile. The city of Raleigh, North Carolina has a population density in 2019 approaching 3,000 people per square mile. It is likely that the actual unit of density is 100 people per square mile. The untransformed variable clearly exhibits positive skew. A natural logarithm transformation can be applied since density is bounded on $(0, \infty]$. Applying the transformation results in a distribution closer to normal. However, thinking about changes in density in terms of 100 people per square mile is tractable. The new model will use the untransformed variable.

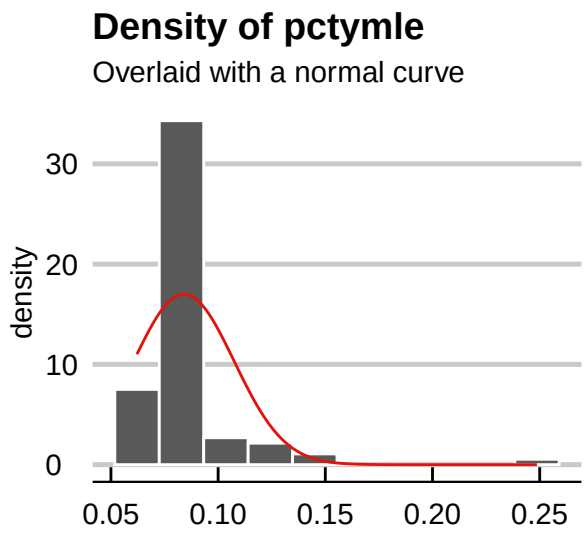
```
density_plot <- econHist('density',
                         crime)
ldensity_plot <- econHist('ldensity',
                          crime)
grid.arrange(density_plot,
             ldensity_plot,
             ncol = 2)
```





The histogram of `pctymle` also shows a positive skew. However, to maintain interpretability the variable is left untransformed.

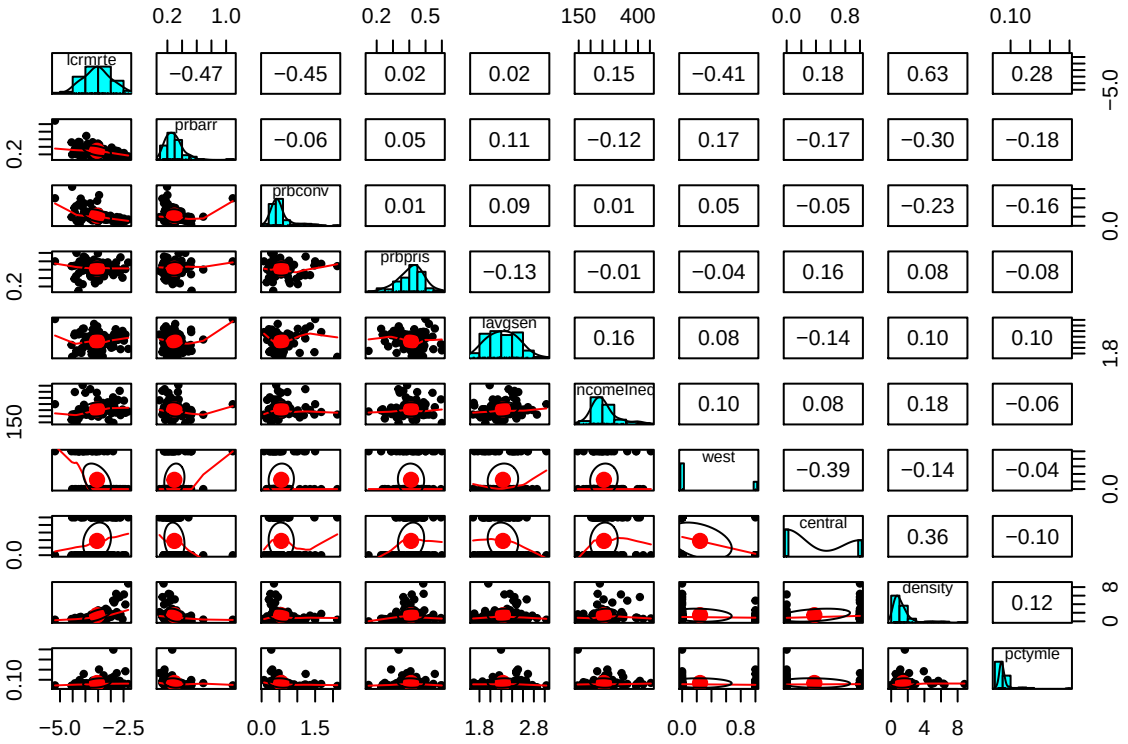
```
pctymle_plot <- econHist('pctymle',
                         crime)
grid.arrange(pctymle_plot,
             ncol = 2)
```



Multivariate analysis

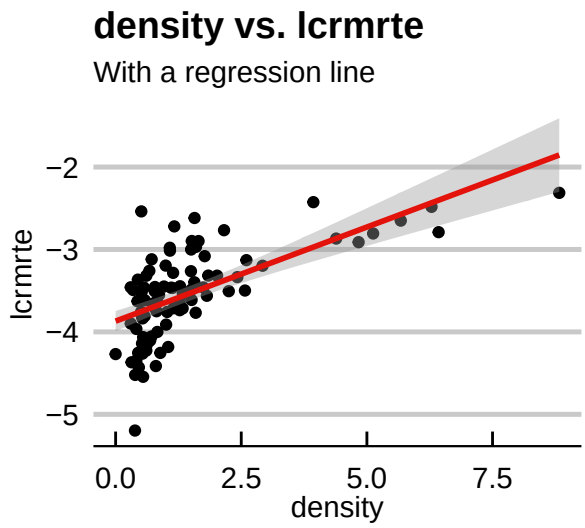
A scatterplot matrix gives a sense of the relationships between the new variables and those in the base model.

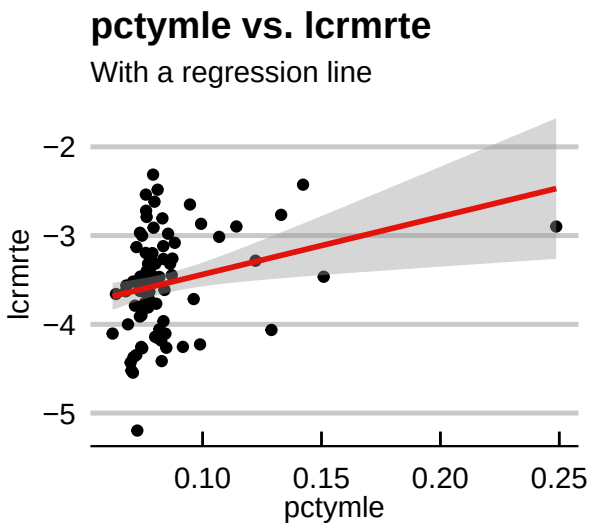
```
crime_2 <- crime %>% select('lcrmrte', 'prbarr', 'prbconv', 'prbpris',
                            'lavgsen', 'incomeIneq', 'west', 'central',
                            'density', 'pctymle')
pairs.panels(crime_2)
```



Several of the new explanatory variables are correlations with `lmcmrte`. In particular, the relationships between `lmcrmrte` and the independent variables `density` and `pctymle` appear linear or nearly so. Examining these relationships with the scatterplot below suggests that the relationship between `pctymle` and `lcrmrte` is likely less strong than that between `density` and `lcrmrte`.

```
density_scatter <- econPlot('density', 'lcrmrte', crime)
pctymle_scatter <- econPlot('pctymle', 'lcrmrte', crime)
grid.arrange(density_scatter, pctymle_scatter, ncol = 2)
```





Fitting the model

The new population model follows.

$$log(crmrte) = \beta_0 + \beta_1 prbarr + ... + \beta_6 log(incomeIneq) + \beta_7 pctymle + \beta_8 density + \beta_9 west + \beta_{10} central + u$$

We fit the model using OLS and obtain the following coefficients and associated robust standard errors.

```
fit_2 <- lm(formula = lcrmrte ~ prbarr + prbconv +
              prbpris + lavgsen + lpolpc +
              lincomeIneq + pctymle + density +
              west + central, data = crime)
fit_2r2 <- summary(fit_2)$r.squared
fit_2Coeftest <- coeftest(fit_2, vcov = vcovHC)
fit_2Coeftest %>% coefTestTable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.095	1.756	-0.054	0.957
prbarr	-1.634	0.303	-5.389	0.000
prbconv	-0.577	0.095	-6.067	0.000
prbpris	0.168	0.455	0.370	0.712
lavgsen	-0.123	0.130	-0.948	0.346
lpolpc	0.462	0.121	3.817	0.000
lincomeIneq	0.090	0.185	0.486	0.628
pctymle	0.879	1.487	0.591	0.556
density	0.119	0.025	4.710	0.000
west	-0.486	0.085	-5.749	0.000
central	-0.209	0.064	-3.281	0.002

The new model's R^2 is `0.799` which is substantially higher than the base model's R^2 of `0.61`. We interpret this to mean that the new model's independent variables are more able to explain the variance in the model than were the base model's.

Interpreting the coefficients

β_1 has a smaller effect than in the base model. In the new model, a 1% increase in `prbarr` is associated with a `-1.634`% change in the crime rate.

β_2 also has a smaller effect Now, we associate a 1% increase in `prbconv` with a `-0.577`% change in the crime rate.

β_5 has a much smaller effect in the new model. A 1% increase in the number of police per capital is associated with a `0.462`% change in the crime rate.

β_8 is interpreted as a unit change in density (again, thought to mean 100 people per square mile, though this is unclear for the reasons stated above) being associated with a change in the crime rate of `11.909`%.

β_9 implies that western counties experience a parallel shift in the crime rate of `-48.617`% when compared to eastern counties, the reference category.

β_9 implies that central counties experience a parallel shift in the crime rate of `-20.941`% when compared to the reference category.

The remainder of the coefficients are not statistically different than zero and, therefore, likely have little effect.

Evaluating joint significance

Independent variables of questionable significance seem to be accumulating in the model. These variables are not independently significant, but they may be jointly significant. The hypothesis that `prbpris`, `lavgsen`, `lincomeIneq`, and `pctymle` are not jointly significant can be evaluated with an F-test.

```
linearHypothesis(fit_2,
                 c("lincomeIneq=0", "prbpris=0",
                   "lavgsen=0", "pctymle=0"),
                 white.adjust = "hc1") # Conduct a robust test
```

```
## Linear hypothesis test
##
## Hypothesis:
## lincomeIneq = 0
## prbpris = 0
## lavgsen = 0
## pctymle = 0
##
## Model 1: restricted model
## Model 2: lcrmrte ~ prbarr + prbconv + prbpris + lavgsen + lpolpc + lincomeIneq +
##     pctymle + density + west + central
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
## 1     83
## 2     79  4 0.51   0.73
```

We fail to reject the null hypothesis that the variables under investigation are not jointly significant. In the interest of parsimony, we remove these variables from the model and compare the results.

```
fit_2a <- lm(formula = lcrmrte ~ prbarr + prbconv +
               lpolpc + density + west + central,
             data = crime)
fit_2ar2 <- summary(fit_2a)$r.squared
fit_2aCoeftest <- coeftest(fit_2a, vcov = vcovHC)
fit_2aCoeftest %>% coefTestTable()
```

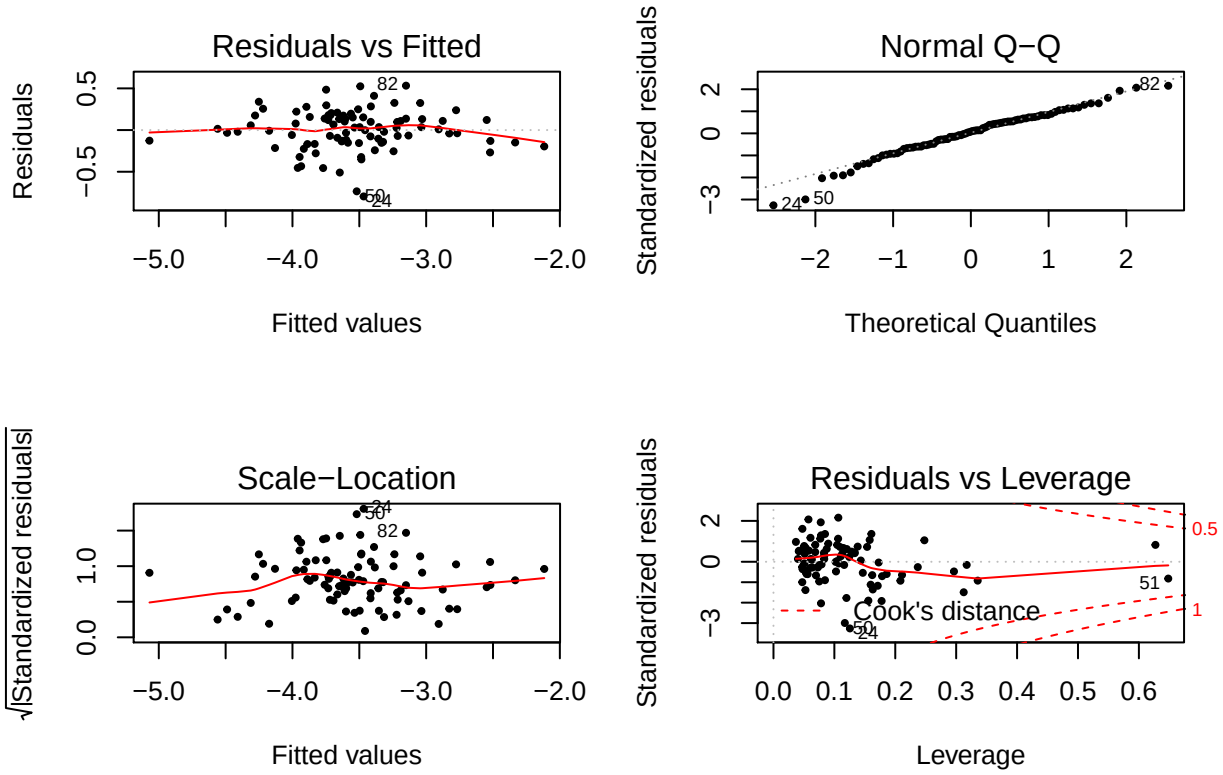
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.260	0.591	0.441	0.661
prbarr	-1.699	0.257	-6.606	0.000
prbconv	-0.596	0.085	-6.968	0.000
lpolpc	0.457	0.086	5.303	0.000
density	0.118	0.023	5.171	0.000
west	-0.482	0.079	-6.094	0.000
central	-0.196	0.062	-3.154	0.002

The parsimonious model has a similar R^2 value of `0.794` and nearly unchanged coefficients on the independent variables.

Highlights of evaluation of the OLS assumptions

Examining the diagnostic plots for the parsimonious model reveals that the zero conditional mean assumption appears to have been met. Slight deviations from zero mean exist for large fitted values, but those deviations are small and are in an area of very sparse data. The scale-location plot seems to show a relatively constant increase which may indicate heteroscedasticity. A Breusch-Pagan test ($p =$ `0.177`), however, suggests that we cannot reject the null hypothesis of homoscedasticity. In any event, robust estimators address the possibility of heteroscedasticity.

```
par(mfrow = c(2,2))
plot(fit_2, pch = 19, cex = 0.5)
```



A Shapiro-Wilk test of the residuals suggests that we cannot reject the null hypothesis of normality, but the results are borderline at $p = 0.07$. Fortunately we can rely on the large sample size and move forward.

Implications

The second model has a high value of R^2 and confirms the direction, if not the magnitude, of the most interesting findings in the base model. Since the OLS assumptions are met, the second model can be confidently applied to policy decisions. The key variables that stand to impact the crime rate and that can be affected through policy are the probability of arrest and the probability of conviction. The coefficients on these variables suggest the value in providing police and prosecutors with the tools they need to effect arrests and secure convictions.

3.3 A Larger Model

The final model will consist of the bulk of the remaining available explanatory variables. The intent of this final modeling effort is to demonstrate the robustness of the results achieved in the preceding two models.

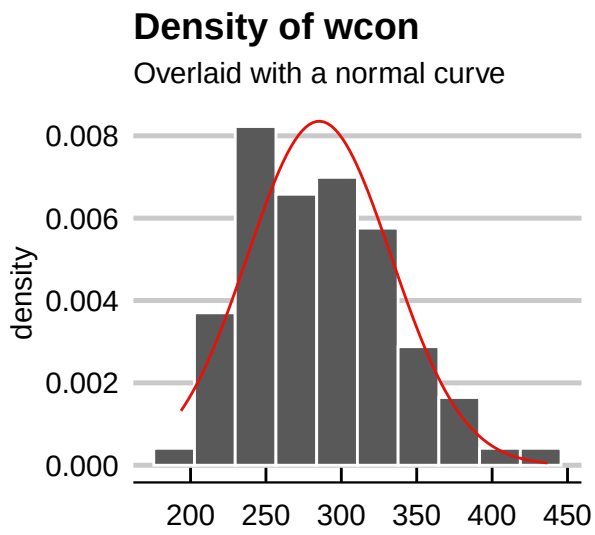
This model will add variables related to:

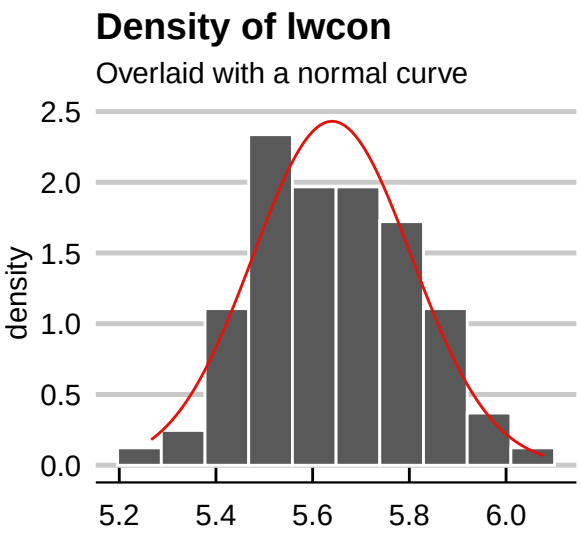
1. Income, in the form of the nine wage variables with the exception of `wser` due to the large outlier noted in the EDA above.
2. Demographics in the form of data on the racial makeup of the counties reported in `pctmin80`.
3. The severity of crime reported in `mix`.
4. Relative affluence proxied by tax revenue and reported in `taxpc`.

Univariate analysis of the added explanatory variables

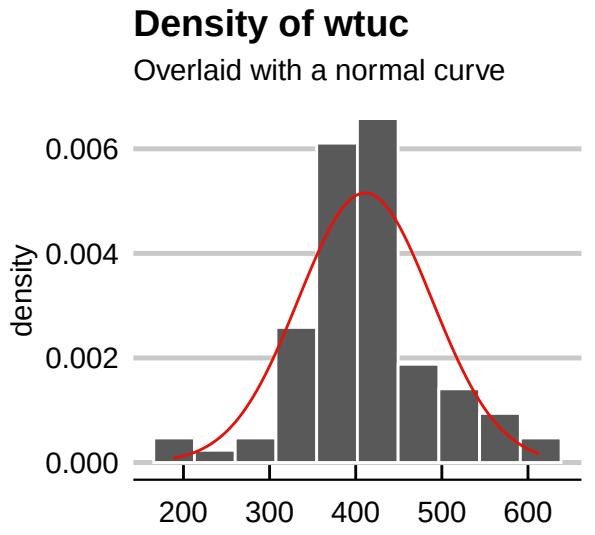
The histograms of the nine wage variables all show a positive skew. A logarithmic transformation makes the variables more interpretable and mitigates the skew. This transformation is valid as wages exist on the range $(0, \infty]$.

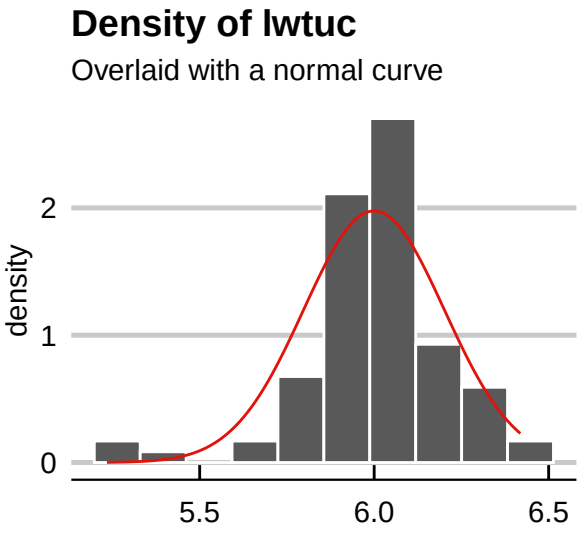
```
wcon_plot  <- econHist('wcon', crime)
lwcon_plot <- econHist('lwcon', crime)
grid.arrange(wcon_plot, lwcon_plot,
             ncol = 2)
```



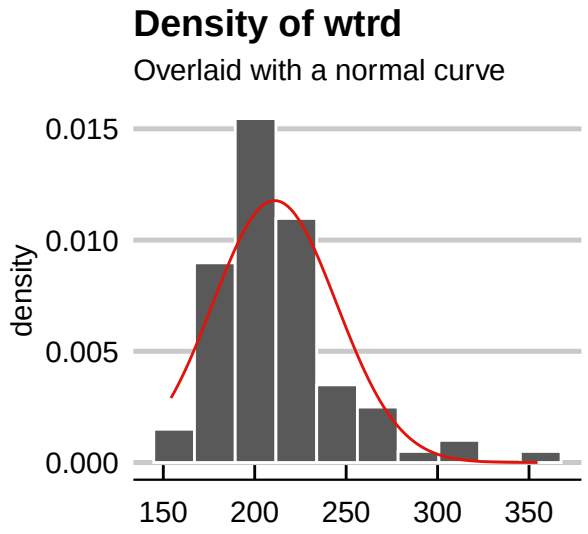


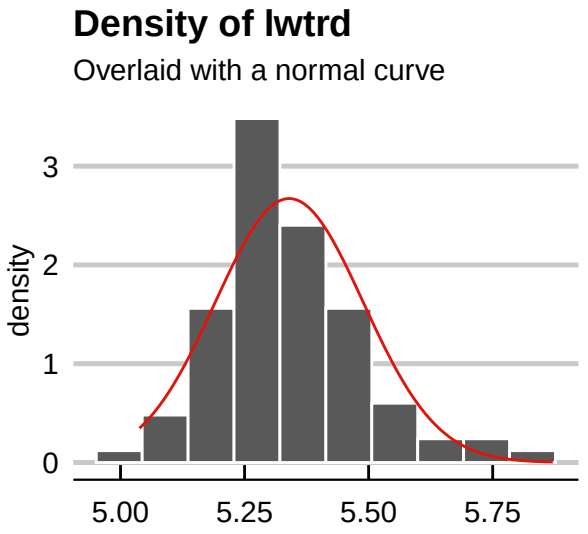

```
wtuc_plot  <- econHist('wtuc', crime)
lwtuc_plot <- econHist('lwtuc', crime)
grid.arrange(wtuc_plot, lwtuc_plot,
             ncol = 2)
```



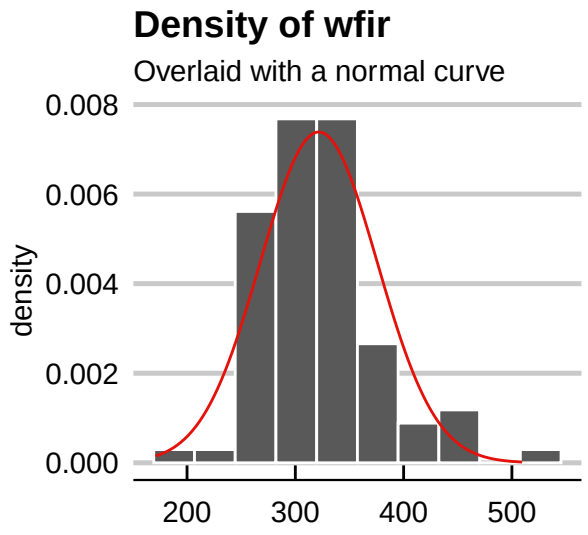


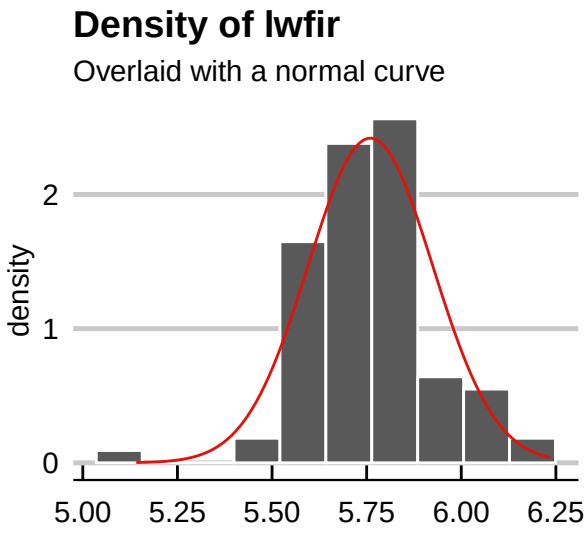
```
wtrd_plot  <- econHist('wtrd', crime)
lwtrd_plot <- econHist('lwtrd', crime)
grid.arrange(wtrd_plot, lwtrd_plot,
             ncol = 2)
```



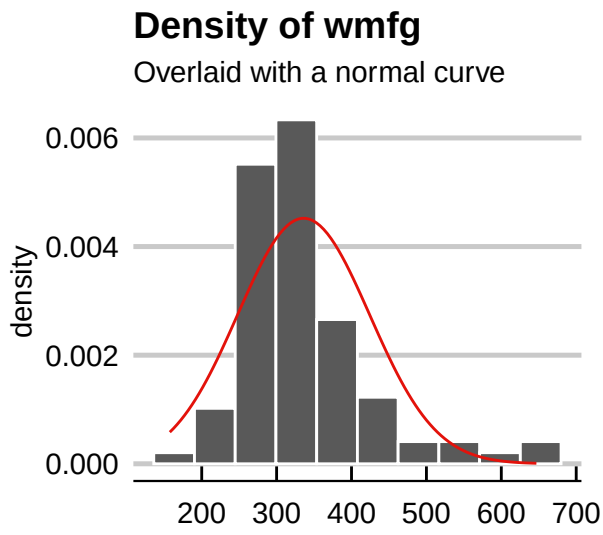


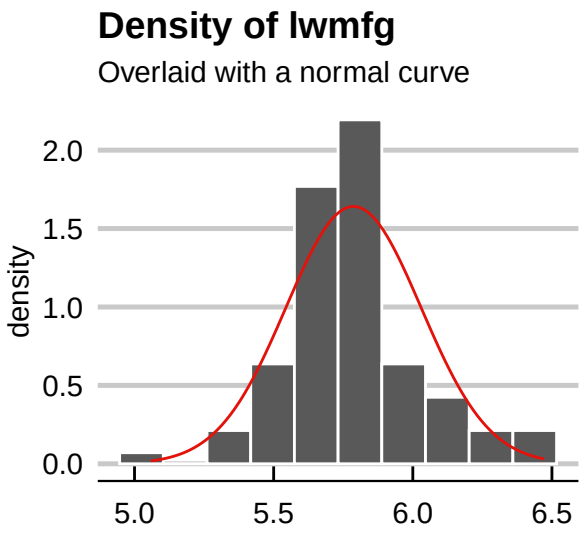

```
wfir_plot  <- econHist('wfir', crime)
lwfir_plot <- econHist('lwfir', crime)
grid.arrange(wfir_plot, lwfir_plot,
             ncol = 2)
```



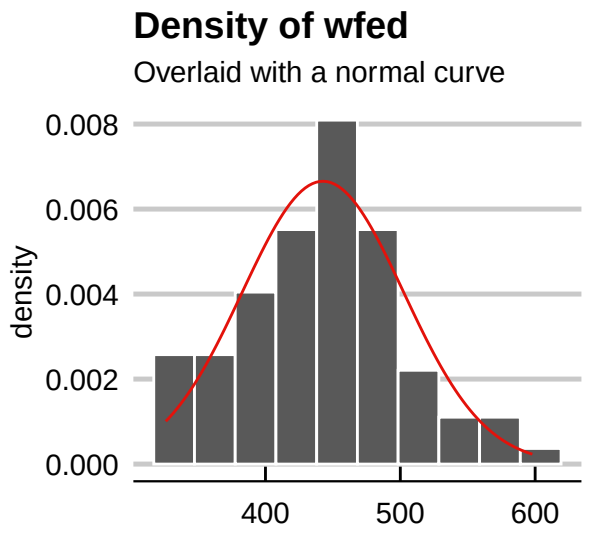


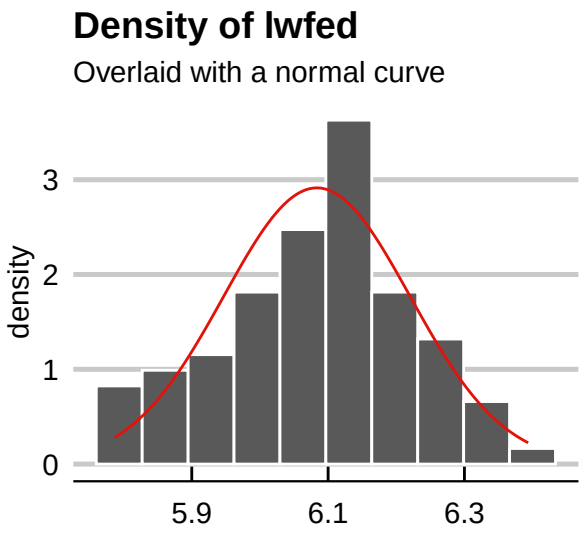

```
wmfg_plot  <- econHist('wmfg', crime)
lwmfg_plot <- econHist('lwmfg', crime)
grid.arrange(wmfg_plot, lwmfg_plot,
             ncol = 2)
```



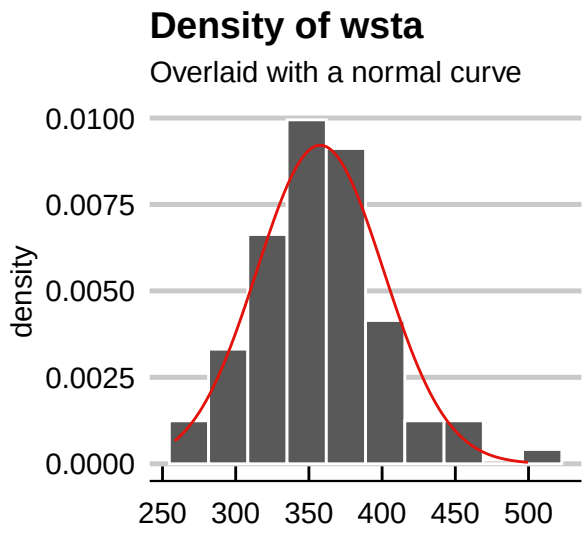


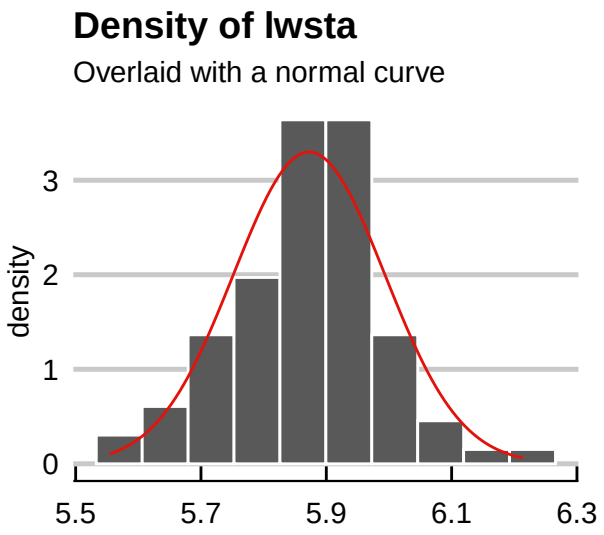

```
wfed_plot  <- econHist('wfed', crime)
lwfed_plot <- econHist('lwfed', crime)
grid.arrange(wfed_plot, lwfed_plot,
             ncol = 2)
```



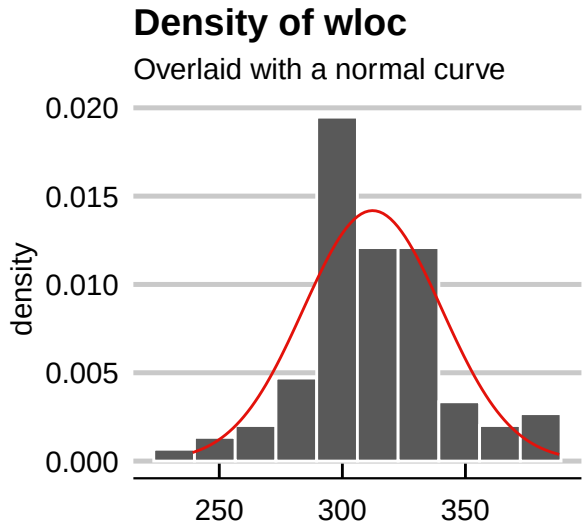


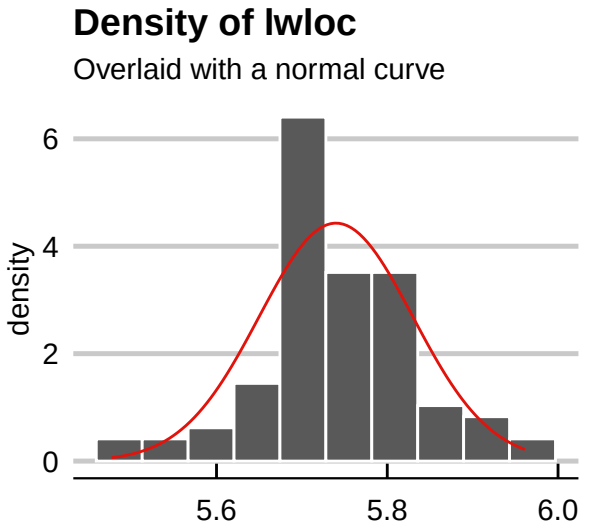

```
wsta_plot  <- econHist('wsta', crime)
lwsta_plot <- econHist('lwsta', crime)
grid.arrange(wsta_plot, lwsta_plot,
             ncol = 2)
```





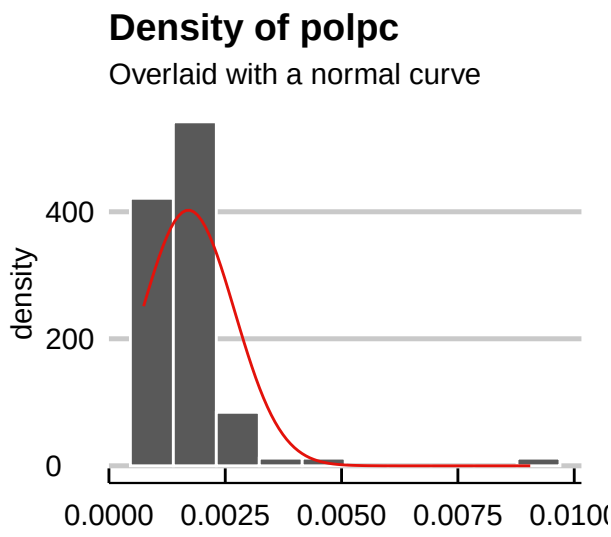

```
wloc_plot  <- econHist('wloc', crime)
lwloc_plot <- econHist('lwloc', crime)
grid.arrange(wloc_plot, lwloc_plot,
             ncol = 2)
```

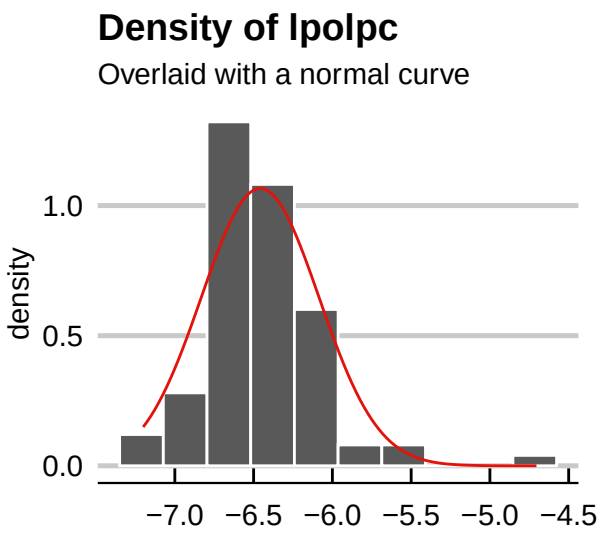




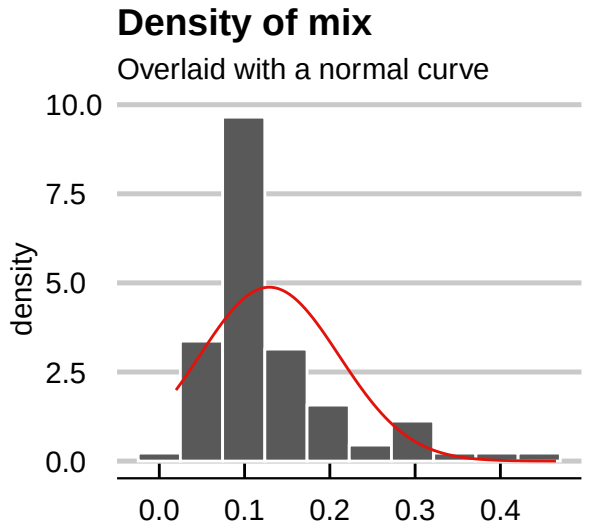
The variables `polpc`, `mix`, and `taxpc` are all skewed. All of these variables on on the range $(0, \infty)$ and are, therefore, amenable to a natural logarithm transformation. Applying the transformation yields the histograms that follow.

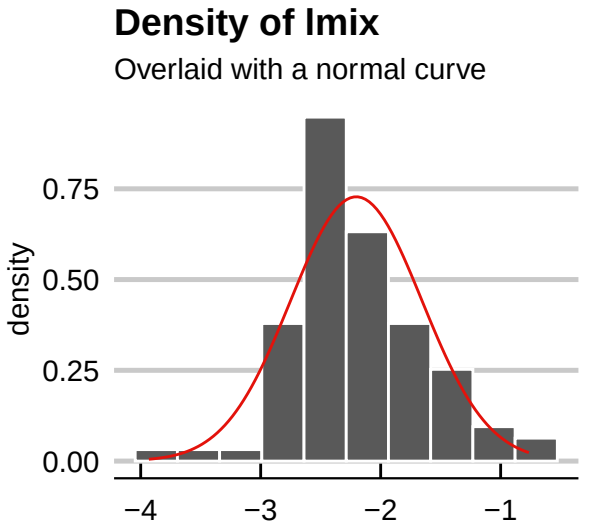
```
polpc_plot  <- econHist('polpc', crime)
lpolpc_plot <- econHist('lpolpc', crime)
grid.arrange(polpc_plot, lpolpc_plot,
             ncol = 2)
```



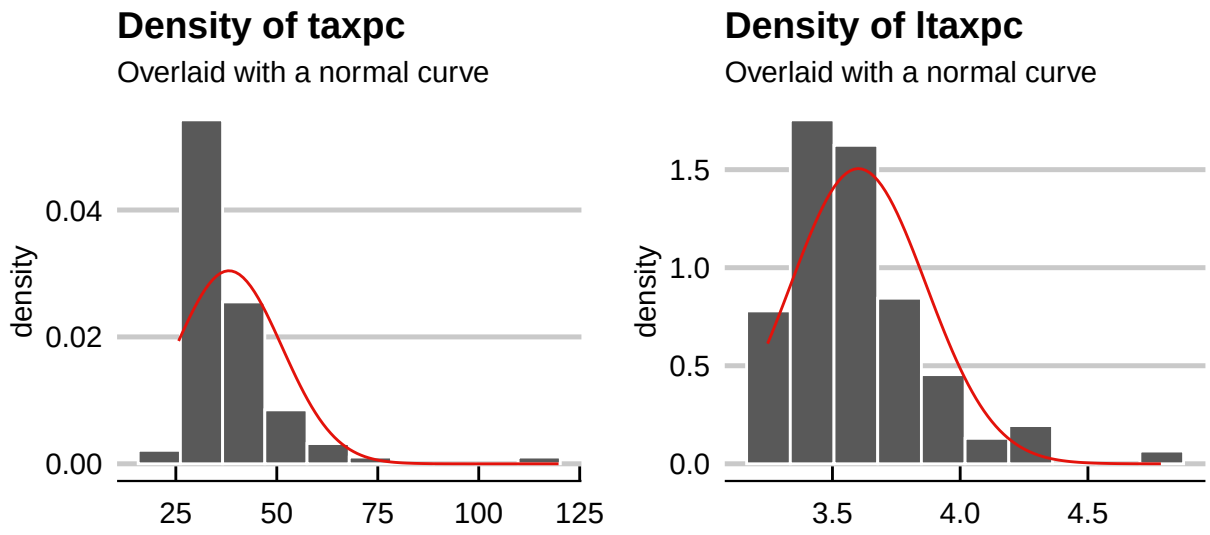



```
mix_plot  <- econHist('mix', crime)
lmix_plot <- econHist('lmix', crime)
grid.arrange(mix_plot, lmix_plot,
             ncol = 2)
```






```
taxpc_plot  <- econHist('taxpc', crime)
ltaxpc_plot <- econHist('ltaxpc', crime)
grid.arrange(taxpc_plot, ltaxpc_plot,
             ncol = 2)
```

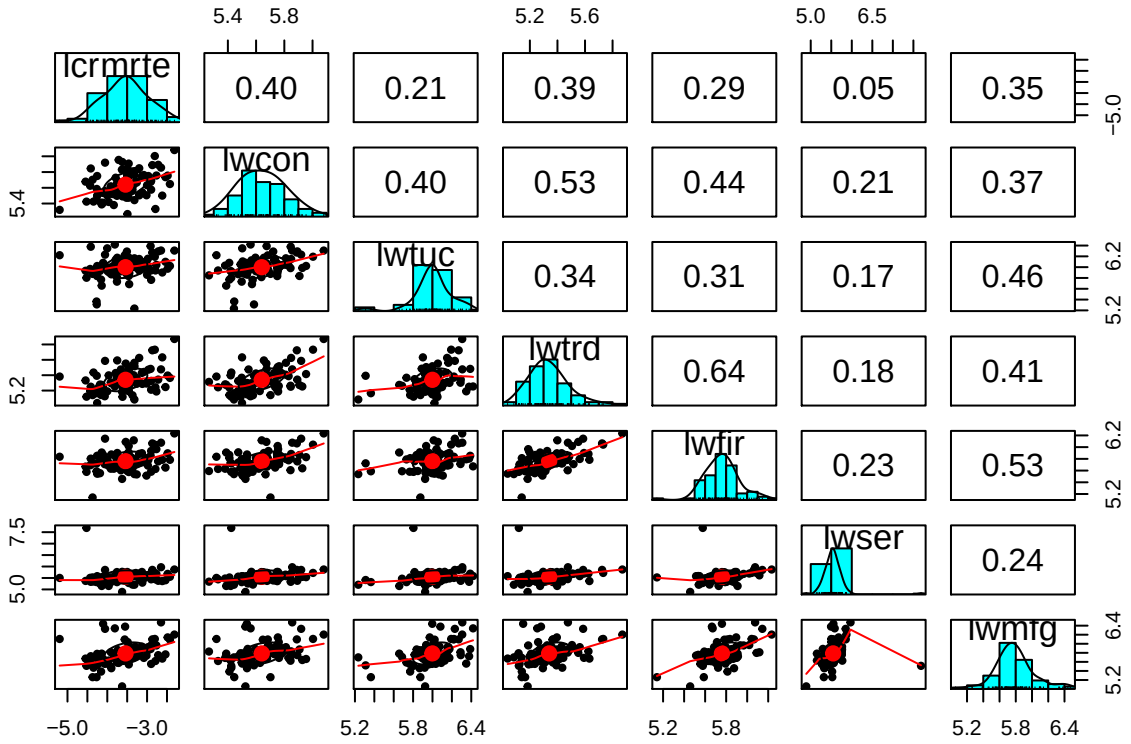


There is no evidence of top- or bottom-coding in the histograms above. The logarithmic transformations largely mitigate the skew in all but the most extreme cases.

Multivariate analysis

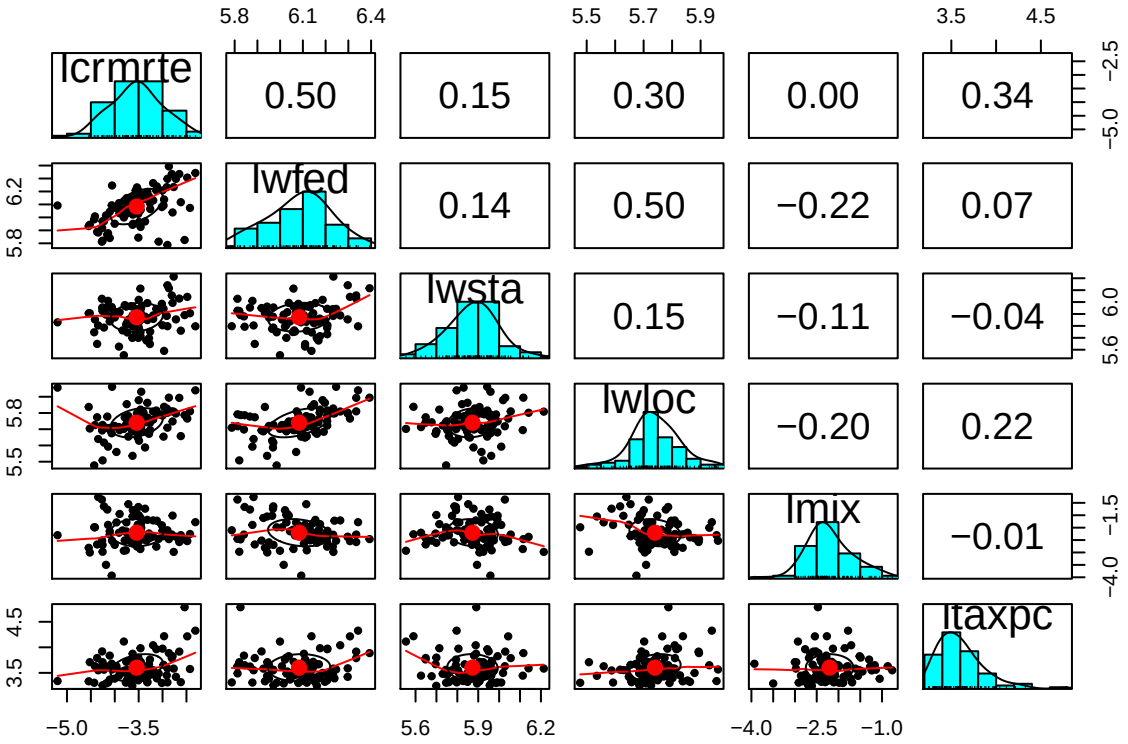
A scatterplot matrix of all of the variables in the final model would be unwieldy. Rather, constructing scatterplot matrices showing the relationship between the new explanatory variables and the outcome variable, `lcrmrte`, may be instructive.

```
crime_3_1 <- crime %>% select('lcrmrte', 'lwcon', 'lwtuc',
                              'lwtrd', 'lwfir', 'lwser',
                              'lwmfg')
pairs.panels(crime_3_1)
```



We note the somewhat surprising correlations between the `lcrmrte` and wages associated with trades typically occupied by uneducated workers including construction, transportation, and manufacturing. Before racing to a conclusion about these correlations, note that the following scatterplot matrix shows that the strongest correlation between `lcrmrte` and a wage variable is with `lwfed`, the average wage of federal workers.

```
crime_3_2 <- crime %>% select('lcrmrte', 'lwfed', 'lwsta',
                              'lwloc', 'lmix', 'ltaxpc')
pairs.panels(crime_3_2)
```



Fitting the model

We fit the model using OLS and compute robust standard errors.

```
fit_3 <- lm(formula = lcrmrte ~ prbarr + prbconv +
              density + west + central + lwcon +
              lwtuc + lwtrd + lwfir + lwmfg +
              lwfed + lwsta + lwloc + lmix +
              lpolpc + ltaxpc,
            data = crime)
fit_3r2 <- summary(fit_3)$r.squared
fit_3Coeftest <- coeftest(fit_3, vcov = vcovHC)
fit_3Coeftest %>% coefTestTable()
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.673	4.411	-0.606	0.546
prbarr	-1.752	0.391	-4.475	0.000
prbconv	-0.618	0.133	-4.655	0.000
density	0.098	0.030	3.233	0.002
west	-0.486	0.079	-6.119	0.000
central	-0.251	0.073	-3.424	0.001
lwcon	0.157	0.250	0.627	0.532
lwtuc	0.076	0.286	0.265	0.792
lwtrd	-0.005	0.443	-0.011	0.991
lwfir	-0.229	0.336	-0.680	0.499
lwmfg	0.034	0.146	0.235	0.815
lwfed	0.686	0.420	1.634	0.106
lwsta	-0.288	0.330	-0.871	0.387
lwloc	0.179	0.657	0.272	0.787
lmix	0.023	0.095	0.240	0.811
lpolpc	0.457	0.163	2.810	0.006
ltaxpc	-0.176	0.223	-0.790	0.432

The final model produces an R^2 of `0.831` which is a small improvement over the second model. More importantly, those variables that were found significant in the previous model remain significant in this model, and their coefficients remain largely unchanged. The results of the previous model are, therefore, robust to the model specification.

Evaluating joint significance

None of the variables added to the previous model are statistically significant. An F-test reveals that the added variables are marginally jointly significant.

```
linearHypothesis(fit_3,
                 c('lwcon=0', 'lwtuc=0', 'lwtrd=0',
                   'lwfir=0', 'lwmfg=0', 'lwfed=0',
                   'lwsta=0', 'lwloc=0', 'lmix=0',
                   'ltaxpc=0'), white.adjust = "hc1")
```

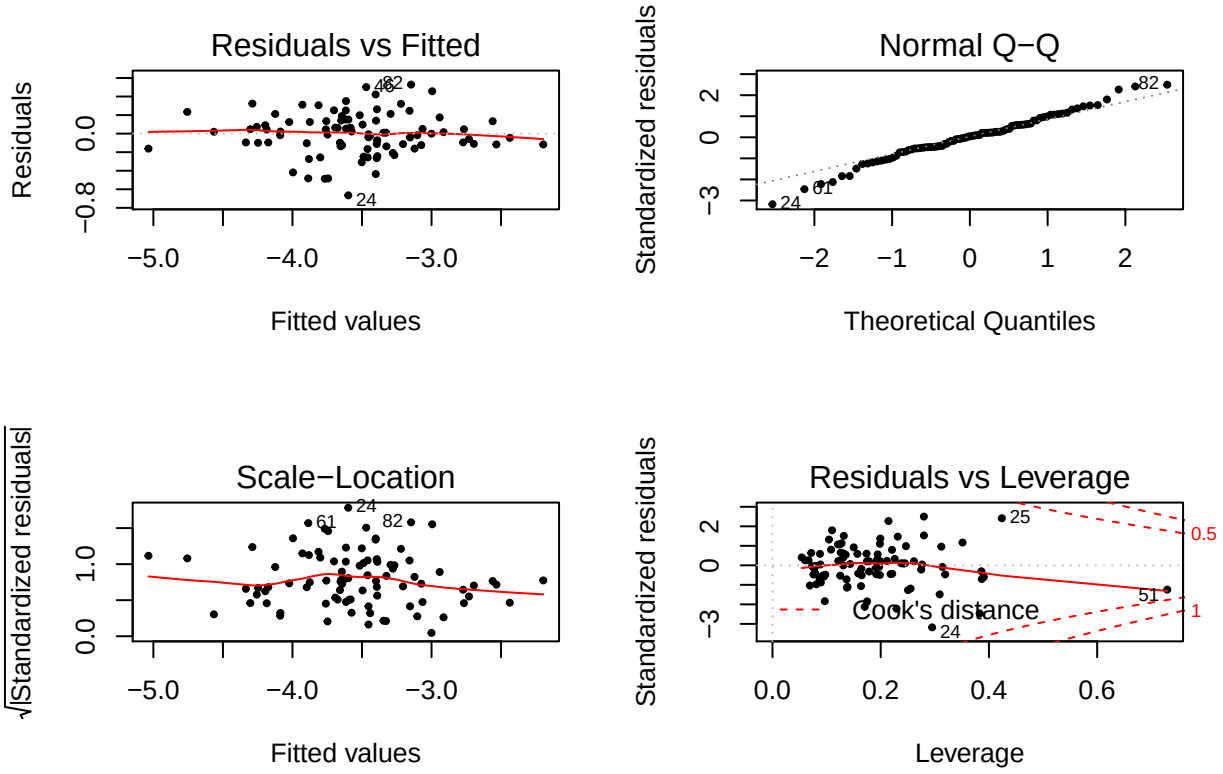
```
## Linear hypothesis test
##
## Hypothesis:
## lwcon = 0
## lwtuc = 0
## lwtrd = 0
## lwfir = 0
## lwmfg = 0
## lwfed = 0
## lwsta = 0
## lwloc = 0
## lmix = 0
## ltaxpc = 0
##
## Model 1: restricted model
## Model 2: lcrmrte ~ prbarr + prbconv + density + west + central + lwcon +
```

```
##     lwtuc + lwtrd + lwfir + lwmfg + lwfed + lwsta + lwloc + lmix +
##     lpolpc + ltaxpc
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df    F Pr(>F)
## 1     83
## 2     73 10 1.94  0.053 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Highlights of evaluating the OLS assumptions

The diagnostic plots reveal no significant deviations from the OLS assumptions. The concern raised during the construction of the base model regarding the effect of spatial autocorrelation on MLR.2 remains. Heteroscedasticity is indicated in the scale-location plot and through a Breusch-Pagan test with $p = 0.001$. Again, we address heteroscedasticity by using robust estimators. MLR.6 remains guaranteed by virtue of the large sample size despite indications that the residuals are not normally distributed.

```
par(mfrow = c(2,2))
plot(fit_3, pch = 19, cex = 0.5)
```



Implications

As stated above the primary purpose of this third model is to determine which variables of interest are robust to changes in the model specification. The argument that the crime rate may be affected by putting policies

in place supporting more aggressive policing and better investigations in support of successful prosecutions is buttressed by these results.

3.4 A Thorough Evaluation of the OLS Assumptions for the Second Model

MLR.1 - Linear in parameters. This assumption is met simply by virtue of the model specification. Our model is of the form $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_j x_j + \hat{u}$.

MLR.2 - Random sampling. We do not have the tools to assess whether this assumption is met. As stated in the discussion of the base model, spatial autocorrelation may be an issue, but this is difficult to assess without understanding the spatial arrangement of the counties under consideration.

MLR.3 - No perfect multicollinearity. R's `lm` function would throw an error if any of the independent variables were perfect linear combinations of one another. Since this has not occurred, we can safely say that MLR.2 is met.

While there is no evidence that MLR.3 has been violated, there is value in assessing imperfect multicollinearity. This is done by computing the variable inflation factor (VIF) of the variables in the model.

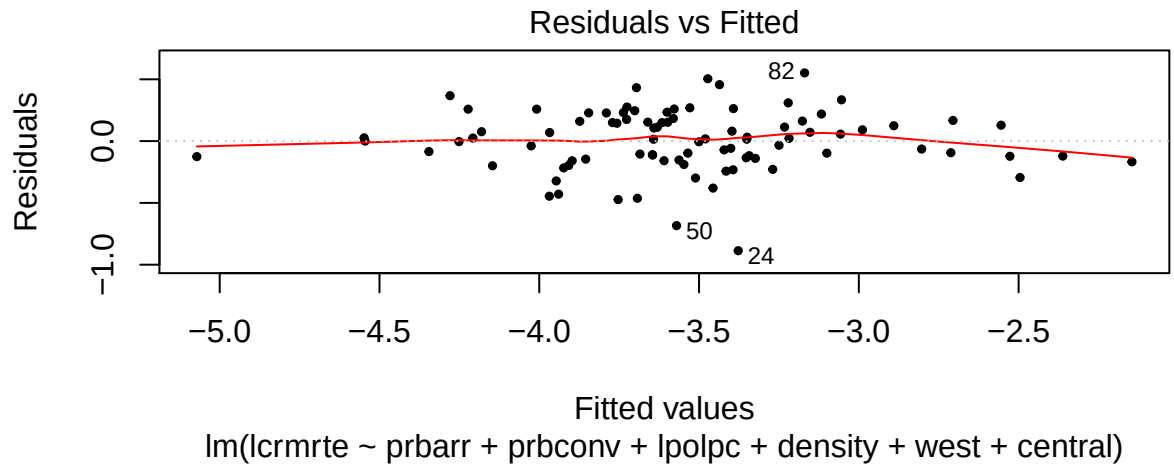
```
vif(fit_2a) %>%
  kable(format = 'latex', booktabs = T) %>%
  kable_styling()
```

	x
prbarr	1.30
prbconv	1.09
lpolpc	1.32
density	1.64
west	1.21
central	1.34

While there is strictly no consensus on what value of VIF is cause for concern, most authors would agree that the values returned for the model do not indicate a problem with the specification.

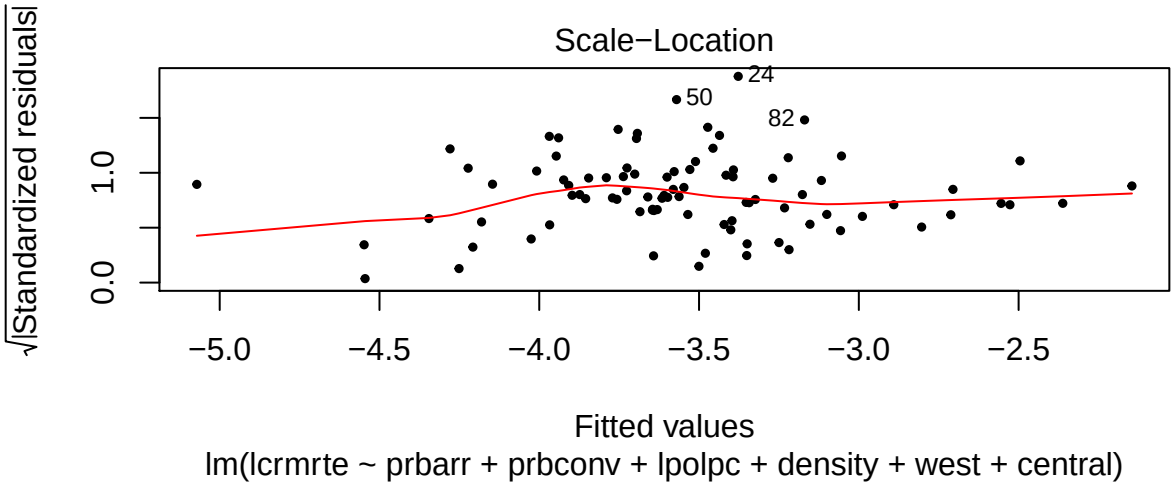
MLR.4 - Zero conditional mean. This assumption can be evaluated graphically. There is little evidence in the residuals versus fitted plot below that the zero conditional mean assumption has been violated. Excursions from zero are small, irregular, and most prominent in areas of sparse data.

```
plot(fit_2a, which = 1, pch = 19, cex = 0.5)
```



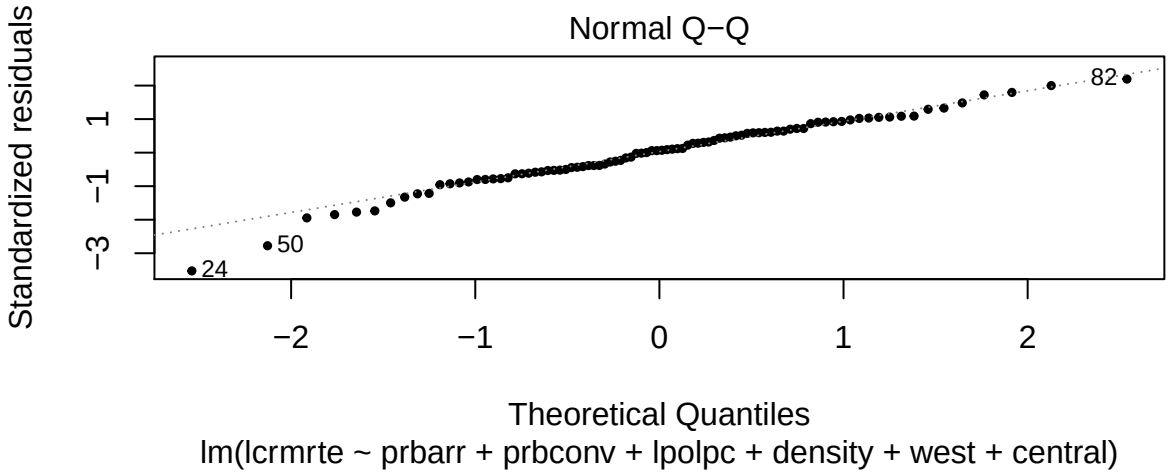
MLR.5 - Homoscedasticity. This is addressed briefly above. The residuals versus fitted plot above shows some indications of heteroscedasticity. Additional evidence to support this is provided by the scale-location plot below. However, a Breusch-Pagan test returns $p = 0.177$. At this value, we fail to reject the null hypothesis of homoscedasticity. To be conservative, however, we have consistently applied robust estimators throughout this study.

```
plot(fit_2a, which = 3, pch = 19, cex = 0.5)
```



MLR.6 - Normality. As has been described above, we suspect that this assumption is violated based on the Q-Q plot below. As was reported above, the results of the Shapiro-Wilks test were marginal and were strictly insufficient to reject the null hypothesis of normality. However, this is largely immaterial as the large sample size allows reliance on OLS asymptotics.

```
plot(fit_2a, which = 2, pch = 19, cex = 0.5)
```



3.5 Regression Table

The results of the three modeling efforts are summarized in the table below.

```
seFit1  <- sqrt(diag(vcovHC(fit_1)))
seFit2a <- sqrt(diag(vcovHC(fit_2a)))
seFit3  <- sqrt(diag(vcovHC(fit_3)))
stargazer(fit_1,
          fit_2a,
          fit_3,
          single.row = TRUE,
          omit.stat = "f",
          se = list(seFit1, seFit2a, seFit3),
          star.cutoffs = c(0.05, 0.01, 0.001),
          type = 'latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, May 07, 2020 - 21:17:03

The practical significance of the results

The value of this study lies in its ability to guide jurisdictions toward actionable policies that have the potential of reducing the crime rate. Consider a notional jurisdiction in the eastern portion of the state with a population of 100,000. The jurisdiction experiences the state average crime rate of roughly 0.034 crimes committed per person annually, or 3,400 crimes per year. The probability of arrest is the state average of 30%, or just over 1,000 arrests per year. Increasing the probability of arrest to 31% results in an associated drop in the crime rate of 1.7% resulting in just under 60 fewer crimes per year. These are not trivial figures, particularly for the potential victims who avoid criminal encounters.

Now consider that the jurisdiction convicts criminals at the state average of 55% or 561 convictions per year. An increase in the conviction rate by 1% is associated with a half percent reduction in the crime rate. Half of a percent, in this case, translates to 17 crimes avoided annually.

Table 1:

	Dependent variable:		
	lcrmrte		
	(1)	(2)	(3)
prbarr	-2.360^{***} (0.369)	-1.700^{***} (0.257)	-1.750^{***} (0.391)
prbconv	-0.734^{***} (0.102)	-0.596^{***} (0.085)	-0.618^{***} (0.133)
prbpris	0.306 (0.739)		
lavgsen	−0.063 (0.191)		
lpolpc	0.626^{***} (0.148)	0.457^{***} (0.086)	0.457^{**} (0.163)
lincomeIneq	−0.042 (0.240)		
ltaxpc			−0.176 (0.223)
density		0.118^{***} (0.023)	0.098^{**} (0.030)
west		-0.482^{***} (0.079)	-0.486^{***} (0.079)
central		-0.196^{**} (0.062)	-0.251^{***} (0.073)
lwcon			0.157 (0.250)
lwtuc			0.076 (0.286)
lwtrd			−0.005 (0.443)
lwfir			−0.229 (0.336)
lwmfg			0.034 (0.146)
lwfed			0.686 (0.420)
lwsta			−0.288 (0.330)
lwloc			0.179 (0.657)
lmix			0.023 (0.095)
Constant	1.850 (2.180)	0.260 (0.591)	−2.670 (4.410)
Observations	90	90	90
R^2	0.610	0.794	0.831
Adjusted R^2	0.582	0.779	0.793
Residual Std. Error	0.355 (df = 83)	0.258 (df = 83)	0.249 (df = 73)
Note:			*p<0.05; **p<0.01; ***p<0.001

3.6 Omitted Variables

We focus our examination of omitted variables on the parsimonious model developed in section 3.2 and evaluated more thoroughly in 3.4. We consider the following omitted variables:

1. **Education**. We assume that in increase in the level of education is associated with a decrease in the crime rate. Further, we assume that a more educated population has a more sophisticated understanding of the criminal justice system and is better equipped to evade law enforcement. Brought before a more educated jury, however, the benefit of education to the criminal in the courtroom attenuates to zero. We expect that the effect of education on the relative size of the police force will be minimal. Assuming that education attracts people, we expect that an increase in the education level will be associated with an increase in population density. Considering the impact of omitting education, we arrive at the following conclusions about the changes in significance of the key explanatory variables.

Variable	β	α	$\alpha\beta$	Significance
`education`	negative			
`prbarr`	negative	negative	positive	decreases
`prbconv`	negative	-	-	-
`lpolpc`	positive	-	-	-
`density`	positive	positive	negative	decreases

2. **Gang activity**. Gang activity is anecdotally associated with an increase in the crime rate. Gangs have a chilling effect on informants and witnesses that reduce the probabilities of arrest and conviction. However, gangs are likely also associated with an increase in violent crimes that result in lengthy prison sentences. Gang activity can certainly be expected to drive an increase in the relative size of the police force. Anecdotal evidence suggests that gang activity may not be associated with a change in density.

Variable	β	α	$\alpha\beta$	Significance
`gangs`	positive			
`prbarr`	negative	negative	negative	increases
`prbconv`	negative	negative	negative	increases
`lpolpc`	positive	positive	positive	increases
`density`	positive	-	-	-

3. **Trust in law enforcement**. We assume that trust in law enforcement is associated with a decrease in the crime rate. Communities that trust law enforcement are probably more likely to cooperate with police and prosecutors, so the significance of the probability of arrest and conviction increase. We expect that trust in the police would be associated with an increase in the size of the force per capita.

Variable	β	α	$\alpha\beta$	Significance
`trust`	negative			
`prbarr`	negative	positive	negative	increases
`prbconv`	negative	positive	negative	increases
`lpolpc`	positive	positive	negative	decreases
`density`	positive	-	-	-

4. **Economic opportunity**. We expect that an increase in economic opportunity will be associated with

a decrease in the crime rate. Economic opportunity will likely shift the character of crime away from violent activity into activity like white collar crime. These crimes may be challenging to investigate. Economic opportunity should tend to increase population density, but we do not expect that it would affect the relative size of the police force.

Variable	β	α	$\alpha\beta$	Significance
`opportunity`	negative			
`prbarr`	negative	negative	positive	decreases
`prbconv`	negative	negative	positive	decreases
`lpolpc`	positive	-	-	-
`density`	positive	positive	negative	decreases

5. **Marijuana legalization**. We consider the impact of omitting the degree to which marijuana is legalized in the jurisdiction. States that have legalized report small increases in traffic-related crimes and other offenses, but these are surely offset by the effect on the crime rate of decriminalization and by reported substitution away from illicit hard drugs and toward marijuana. Arrests for marijuana possession are common and, likely, are easy to make. So we expect that the probability of arrest will decrease with marijuana legalization. Likewise, convictions for marijuana possession and distribution are likely easy to obtain. Legalization will probably decrease the probability of conviction. We do not expect immediate impacts on the size of the police force or the density of a jurisdiction stemming from marijuana legalization.

Variable	β	α	$\alpha\beta$	Significance
`legalization`	negative			
`prbarr`	negative	negative	positive	decreases
`prbconv`	negative	negative	positive	decreases
`lpolpc`	positive	-	-	-
`density`	positive	-	-	-

4.0 Conclusion

Our analysis demonstrates that the effects of aggressive policing and prosecutions are robust and both statistically and practically significant. The party can proceed confidently with developing policy prescriptions around these variables to reduce the crime rate. However, we need to note that an analysis of this sort, no matter how convincing, cannot capture the unique aspects of implementing policy in a political environment. We also caution that dramatic changes to policies as a result of this analysis may lead to significant unintended consequences. We have argued that increasing police aggressiveness may reduce the crime rate. Consider New York's stop and frisk policy as an extreme example. The outcome of the policy on crime remains debatable, but the outcome on civil liberties and equity in New York are dramatic and well documented. Proceed, then, with caution.