W203 Lab 3, Part 1: Reducing Crime

Stephen Holtz
Jon Mease
Hong Yang

Introduction and Research Question

As members of Berkeley Analytica, a political consultancy, we seek to inform political operatives on what policy decisions could be most useful once they take office. We also seek to help inform their campaigns so they can offer voters a sincere, intellectually honest, and meaningful vision for how a candidate or party could change society.

In this project, we are applying a cross section of data from C. Cornwell and W. Trumball (1994), "Estimating the Economic Model of Crime with Panel Data," Review of Economics and Statistics. The primary objective of this project is to inform policy makers of the value of laws and funding decisions viewed through the lens of reducing crime. Specific policy questions include, providing funding for more police officers; setting guidance or requirements for sentencing of criminals; implementing policies that improve the distribution of minorities across neighborhoods; and implementing policies that keep young males occupied and out of trouble. The secondary objective of this project will be to identify other factors that affect crime that policy makers could account for in formulating a strategy to reduce crime.

Initial Data Loading and Cleaning

First we will load and examine the data set.

```
crime_raw <- read.csv("../data/crime_v2.csv")
```

We note that the last 6 rows of the dataset are `NA` in all columns except `prbconv`. According to the code book `prbconv` is a numeric variable representing the probability of conviction, but it has been loaded as a factor due to the presence of a backtick string character present in row 97.

```
tail(crime_raw[,1:6], 10)
```

```
##    county year    crmrte   prbarr     prbconv  prbpris
## 88    193   87 0.0235277 0.266055 0.588859022 0.423423
## 89    193   87 0.0235277 0.266055 0.588859022 0.423423
## 90    195   87 0.0313973 0.201397 1.670519948 0.470588
## 91    197   87 0.0141928 0.207595 1.182929993 0.360825
## 92     NA   NA        NA       NA                   NA
## 93     NA   NA        NA       NA                   NA
## 94     NA   NA        NA       NA                   NA
## 95     NA   NA        NA       NA                   NA
## 96     NA   NA        NA       NA                   NA
## 97     NA   NA        NA       NA           `       NA
```

We remove last 6 rows from `crime_raw` and create a new data frame, `crime`, and convert `prbconv` from a factor into a numeric column.

```
crime <- crime_raw[1:(nrow(crime_raw)-6),]
crime$prbconv <- as.numeric(levels(crime$prbconv)[crime$prbconv])
```

Next we examine the probability variables `prbarr` (The 'probability' of arrest), `prbconv` (The 'probability' of conviction), `prbpris` (The probability of prison sentence.).

```r
summary(crime[,c('prbarr', 'prbconv', 'prbpris')])
```

```
##      prbarr            prbconv            prbpris
##  Min.   :0.09277   Min.   :0.06838   Min.   :0.1500
##  1st Qu.:0.20568   1st Qu.:0.34541   1st Qu.:0.3648
##  Median :0.27095   Median :0.45283   Median :0.4234
##  Mean   :0.29492   Mean   :0.55128   Mean   :0.4108
##  3rd Qu.:0.34438   3rd Qu.:0.58886   3rd Qu.:0.4568
##  Max.   :1.09091   Max.   :2.12121   Max.   :0.6000
```

Here we see that all probability values are non-negative, but `prbarr` and `prbconv` each have values that are greater than one and therefore not valid probabilities. However, the code book states that "the probability of arrest is proxied by the ratio of arrests to offenses" and "the probability of conviction is proxied by the ratio of convictions to arrests". By these calculations, it is plausible that these probability proxy variables will have values larger than one, so we do not omit these observations.

Next, we identify an unreasonably anomalous value for the service industry wage, `wser`, in row 84 for county 185.

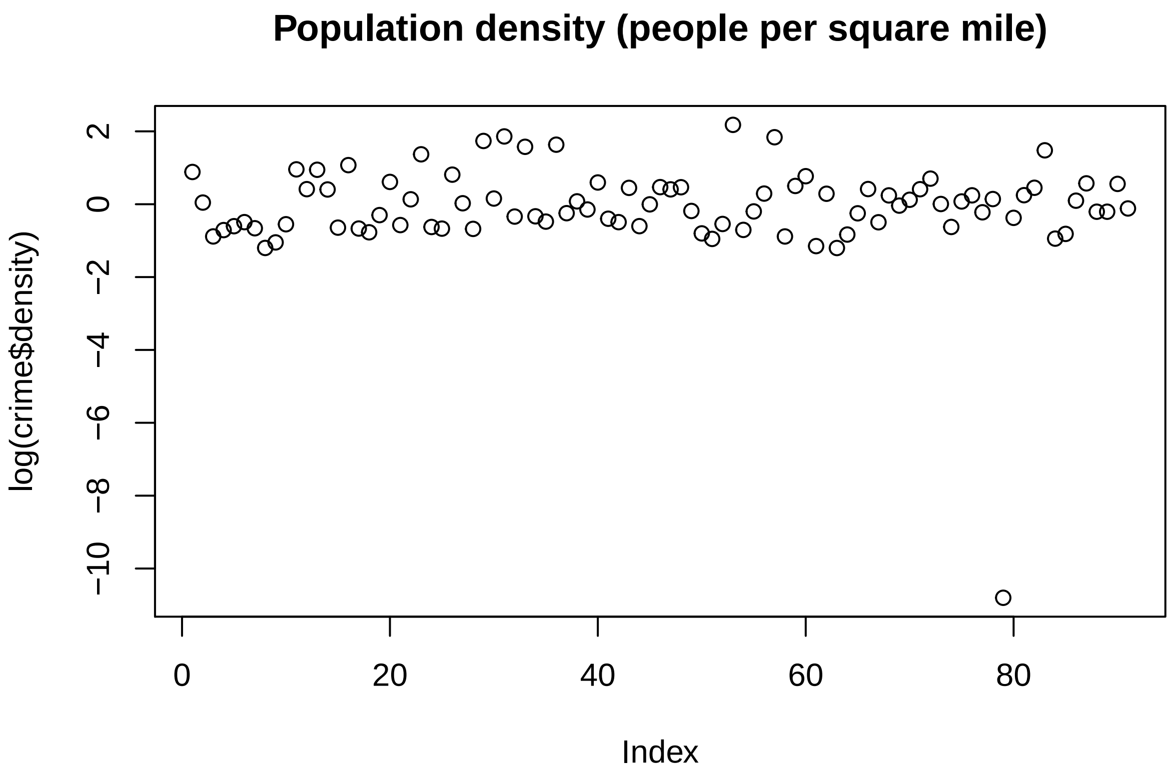
```r
plot(crime$wser, main = 'Service industry weekly wage')
```

This extreme value (2177.0681) is 860% of the median (253.2) and 556% of the second largest wage in the dataset (391.3081). Since the other remaining properties for this observation are all in reasonable ranges, with respect to the other observations in the sample, we will replace the anomalous value with `NA` rather than remove the entire observation.

```r
crime$wser[84] = NA
```

There is a single unreasonably low value of the `density` variable, which we can identify by plotting density on a log scale.

```r
plot(log(crime$density), main = 'Population density (people per square mile)')
```



The 79th observation has a linear scale `density` value of 0.0000203422 people per square mile. The entire state of North Caroline is 53,819 square miles, and if the entire state had this population density there would be only 1.09 persons in the entire state! Therefore a county that is a small fraction of the size of the state that had this density would have less than one person living in it. As above, the remaining variables for this observation all have reasonable values so we once again replace the errant value by `NA` rather than omit the observation.

```r
crime$density[79] = NA
```

The Model Building Process

To investigate our research question the outcome variable of this analysis will be based on the `crmrte` variable.

Outcome variable: `crmrte`

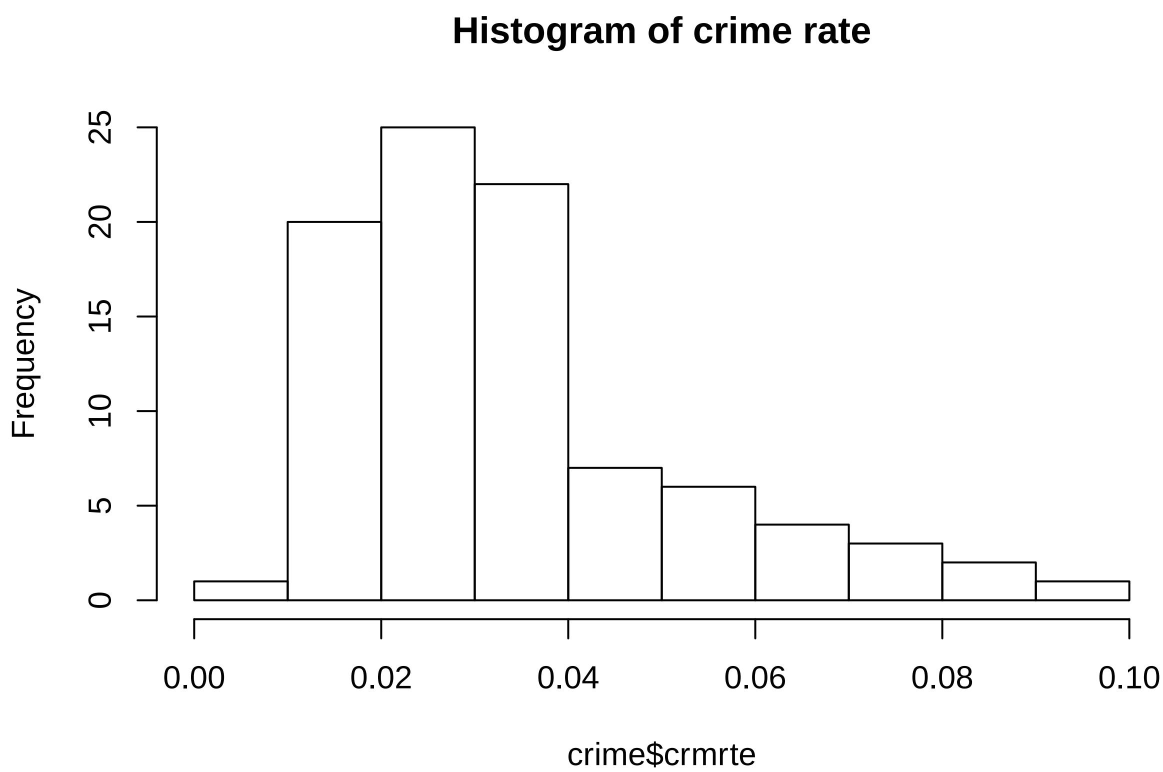
Summarizing `crmrte` we see that the mean is a bit larger than the median indicating a moderate positive skew.

```r
summary(crime$crmrte)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005533 0.020927 0.029986 0.033400 0.039642 0.098966
```

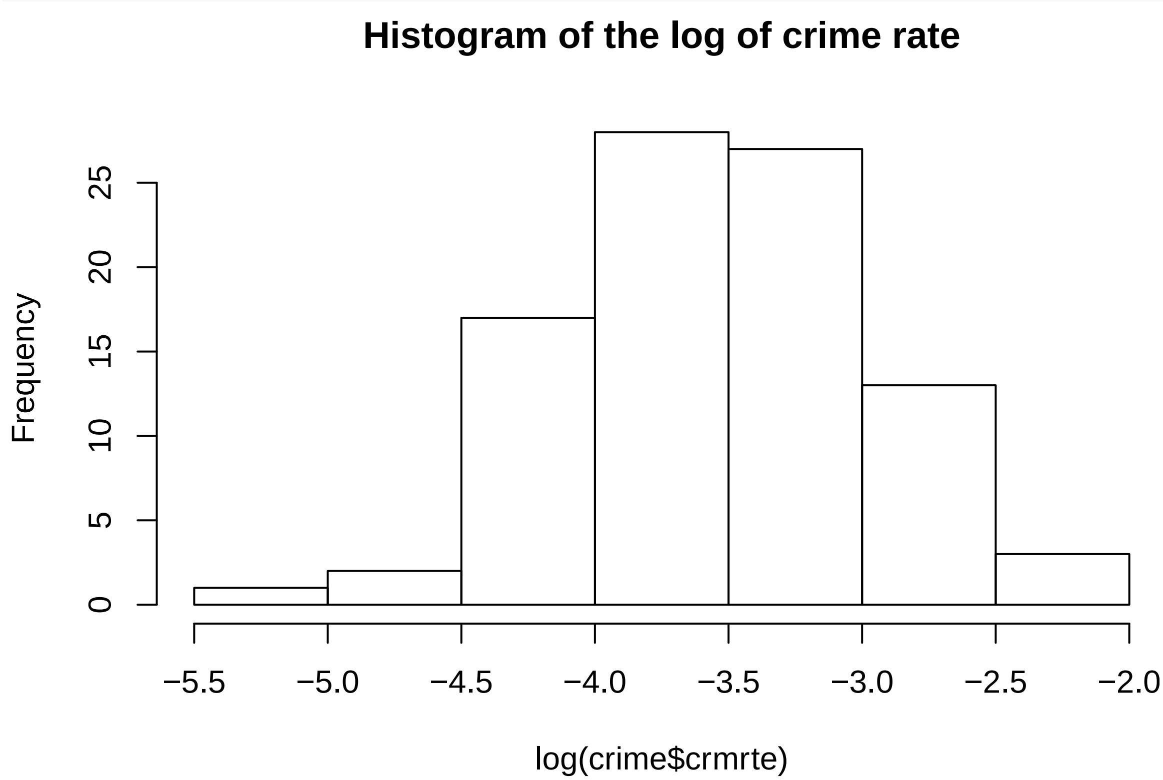
The histogram of `crmrte` confirms the skew and also shows that the distribution of `crmrte` is unimodal.

```r
hist(crime$crmrte, main = "Histogram of crime rate")
```



We note that the skew is almost entirely eliminated by taking the natural log of `crmrte`.

```
hist(log(crime$crmrte), main = "Histogram of the log of crime rate")
```



Based on this, we define our outcome variable as the natural log of `crmrte`. This means that the models we build in this report will describe percentage (not absolute) changes to the crime rate.

Explanatory variables

Because we are working for a political campaign we want to investigate parameters that will inform policy decisions once the political party is in power, as well as the platform that the campaign will use to get elected.

Population density: `density`

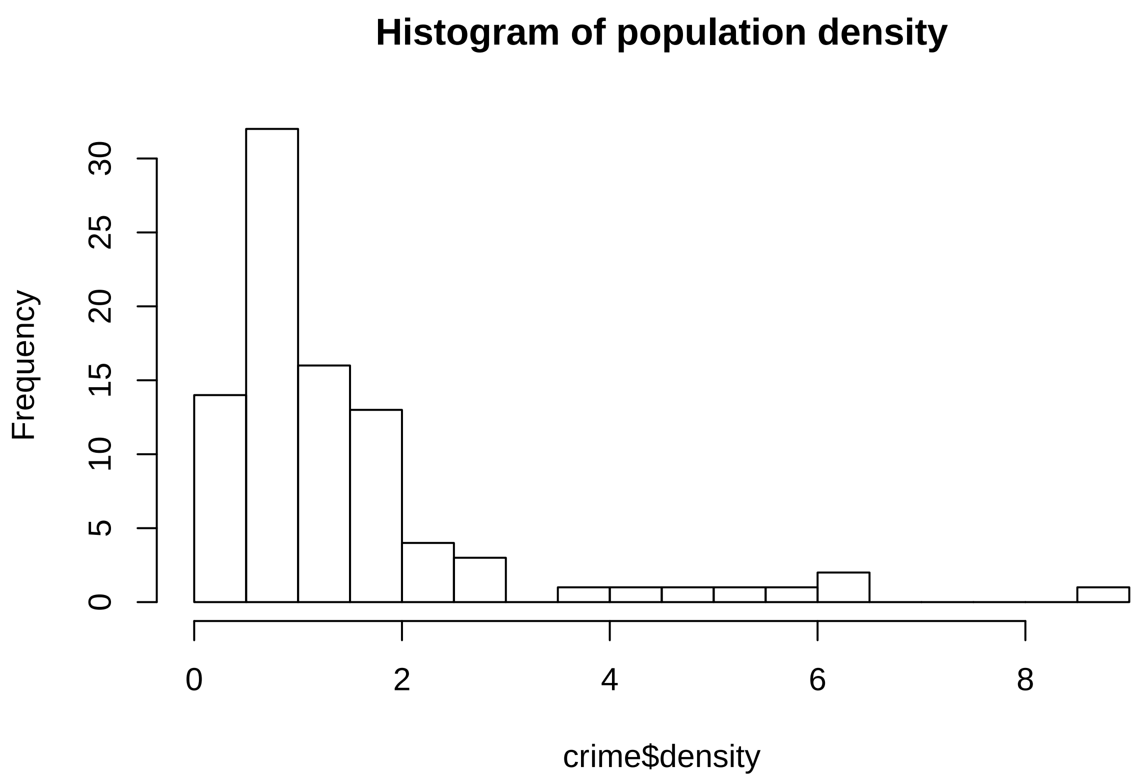
We start `density`, the number of persons per square mile. Here the median is almost 1.5 times larger than the median, indicating a significant positive skew.

```
summary(crime$density)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.3006  0.5519  0.9792  1.4447  1.5693  8.8277       1
```

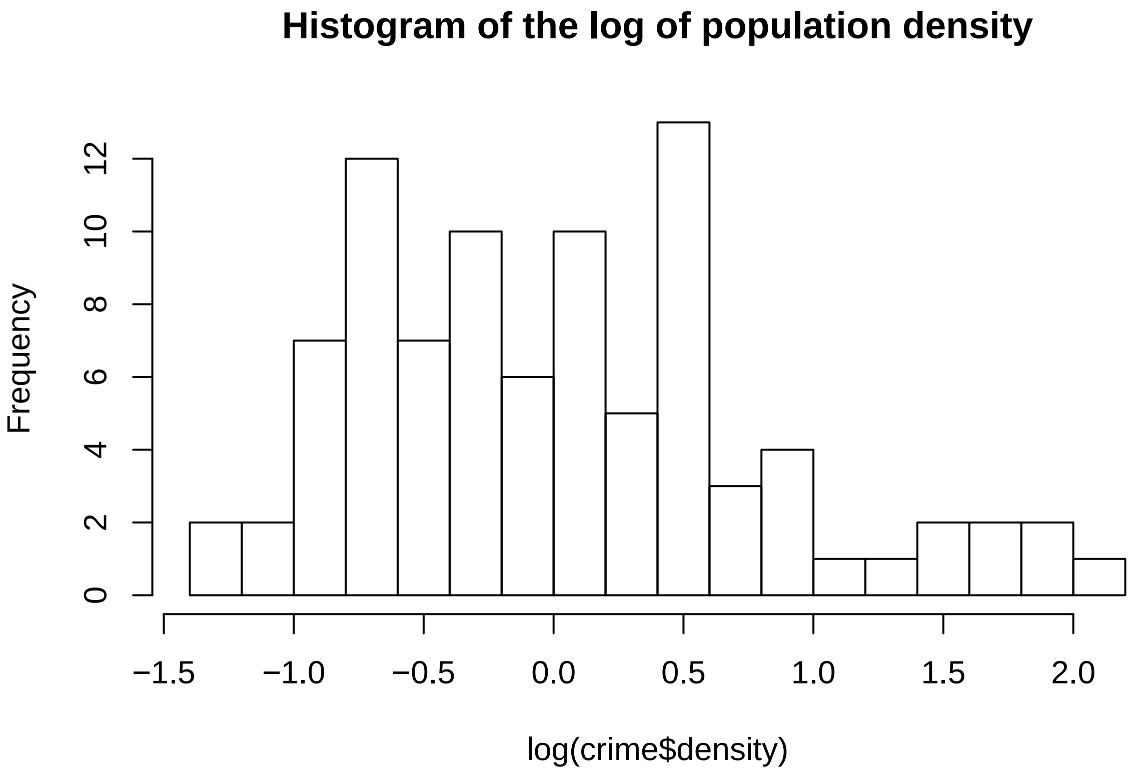
The histogram of `density` confirms the large positive skew and shows that it has a unimodal distribution peaked at around 0.5-1.0 persons per square mile.

```
hist(crime$density, breaks = 15,
     main = "Histogram of population density")
```



We will use the natural log of `density` as our explanatory variable as it is significantly less skewed than `density` itself.

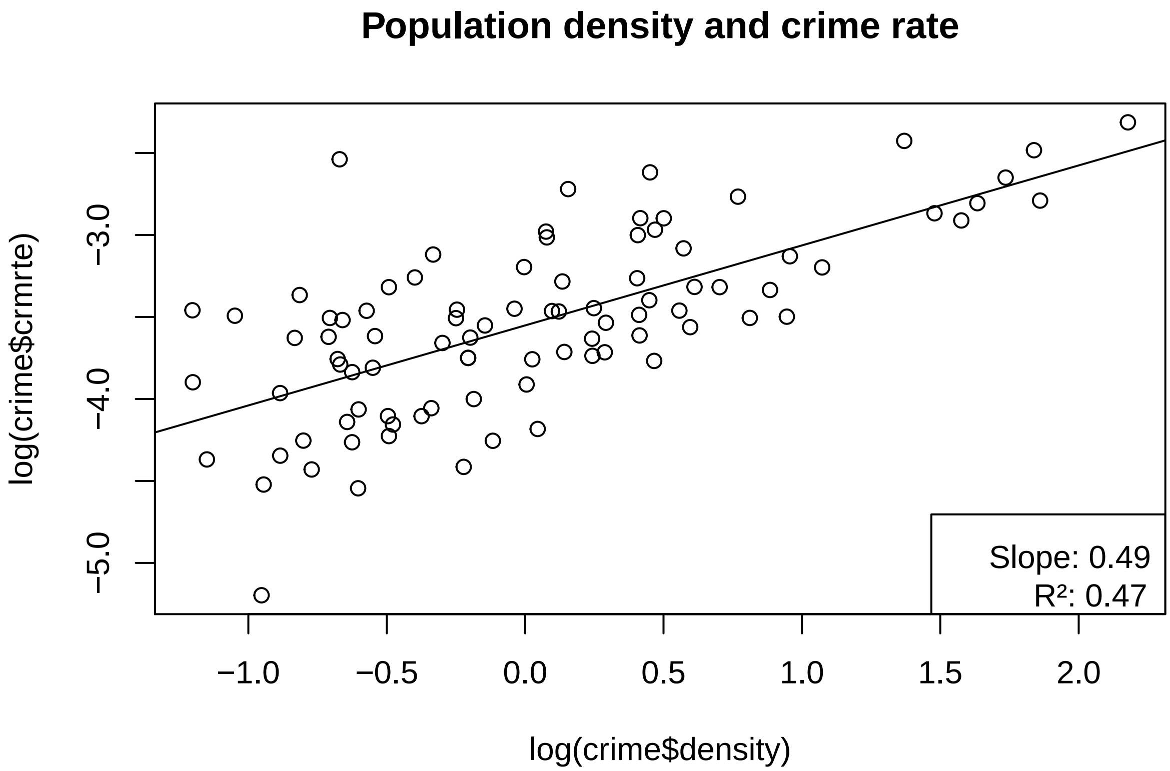
```
hist(log(crime$density), breaks = 15,
     main = "Histogram of the log of population density")
```



In a scatter plot with our outcome variable we note a positive relationship between percentage change in population density and percentage change in crime rate.

```
r2_text <- function(model) {
  return(sprintf('Slope: %.2f\n     R²: %.2f',
                 model$coefficients[[2]], summary(model)$r.squared))
}

model.density <- lm(log(crime$crmrte) ~ log(crime$density))
plot(log(crime$density), log(crime$crmrte),
     main = 'Population density and crime rate')
abline(model.density)
legend('bottomright', legend=r2_text(model.density))
```



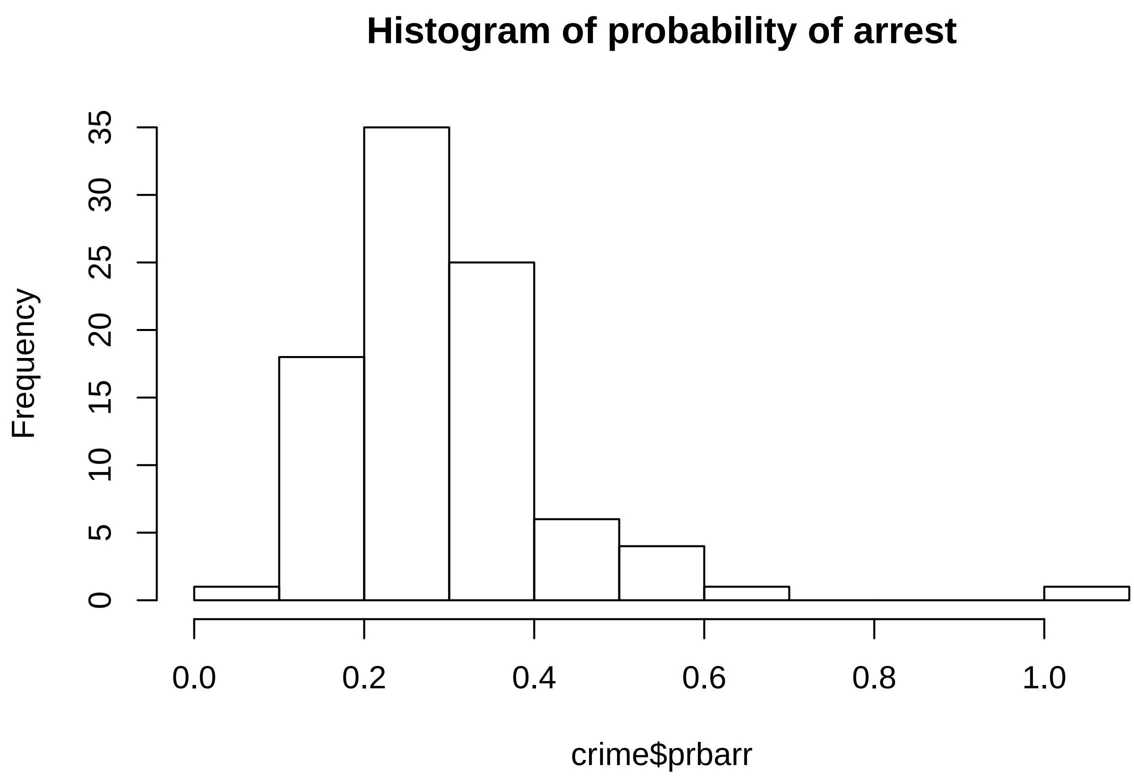
Based on the slope and R^2 statistic of the best-fit line we see that a 1% increase in population density is associated with a ~0.49% increase in the crime rate, and that population density by itself can predict 47% of the variation in crime rate across North Carolina.

Probability of arrest, conviction, and prison sentence: `prbarr`, `prbconv`, and `prbpris`

We believe that probability of arrest is worth investigating as it will inform policy decisions and the platform on whether or not more resources should be put towards increasing the likelihood that a person who commits a crime will be arrested.

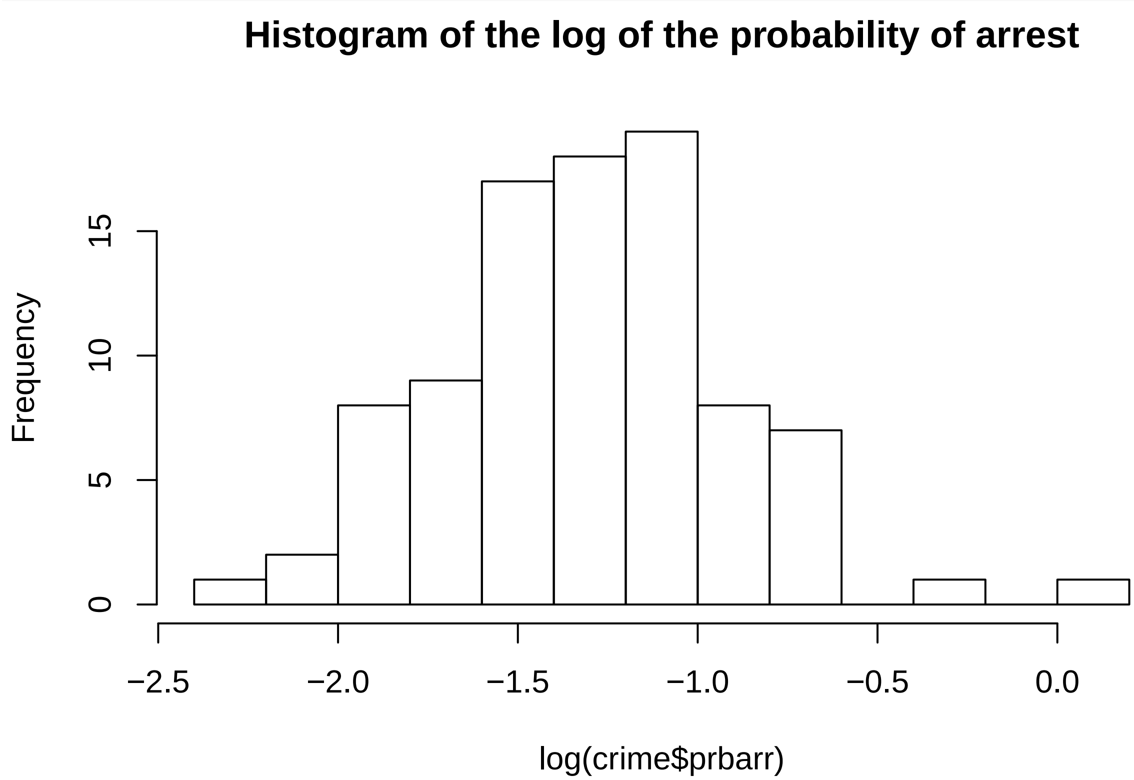
The histogram of the `prbarr` variable shows that the distribution is unimodal with a slight positive skew.

```
hist(crime$prbarr, breaks = 10,
     main = "Histogram of probability of arrest")
```



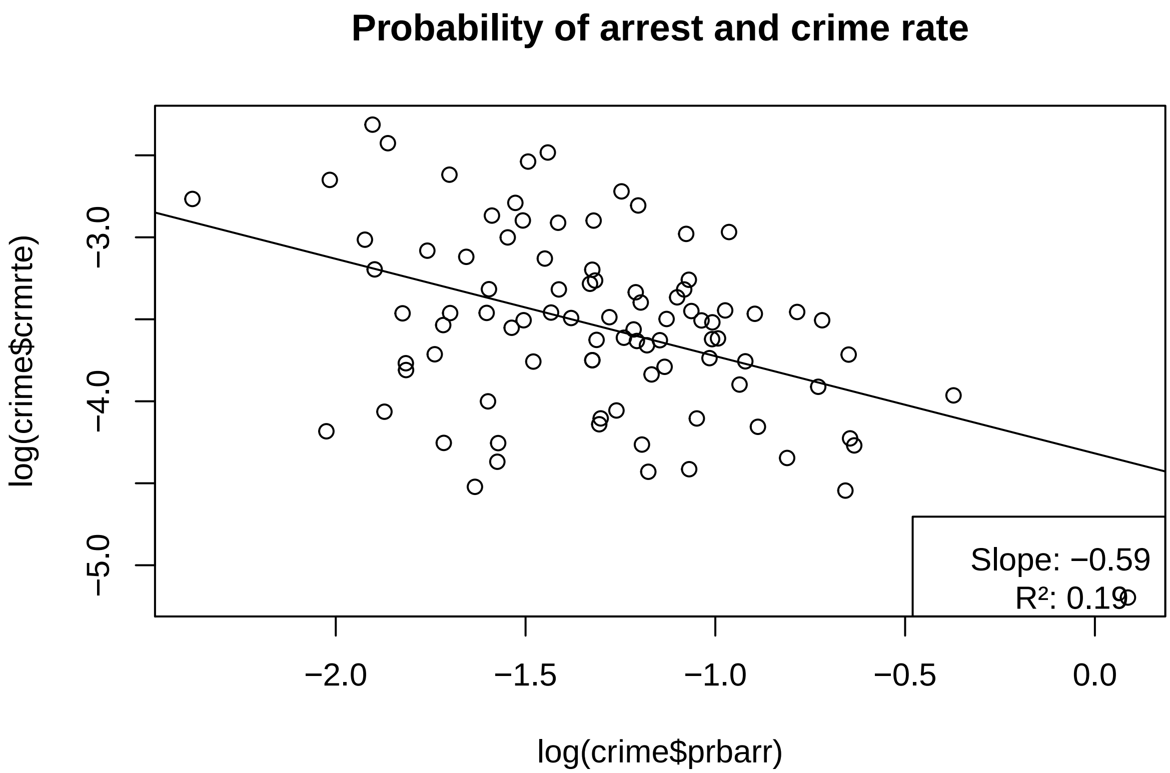
Taking the log transform removes the skew altogether, and so we will use the natural log of `prbarr` as an explanatory variable.

```
hist(log(crime$prbarr), breaks = 10,
     main = "Histogram of the log of the probability of arrest")
```



In a scatter plot with our outcome variable we note a negative relationship between percentage change in probability of arrest and percentage change in crime rate.

```
model.prbarr <- lm(log(crime$crmrte) ~ log(crime$prbarr))
plot(log(crime$prbarr), log(crime$crmrte),
     main = 'Probability of arrest and crime rate')
abline(model.prbarr)
legend('bottomright', legend=r2_text(model.prbarr))
```



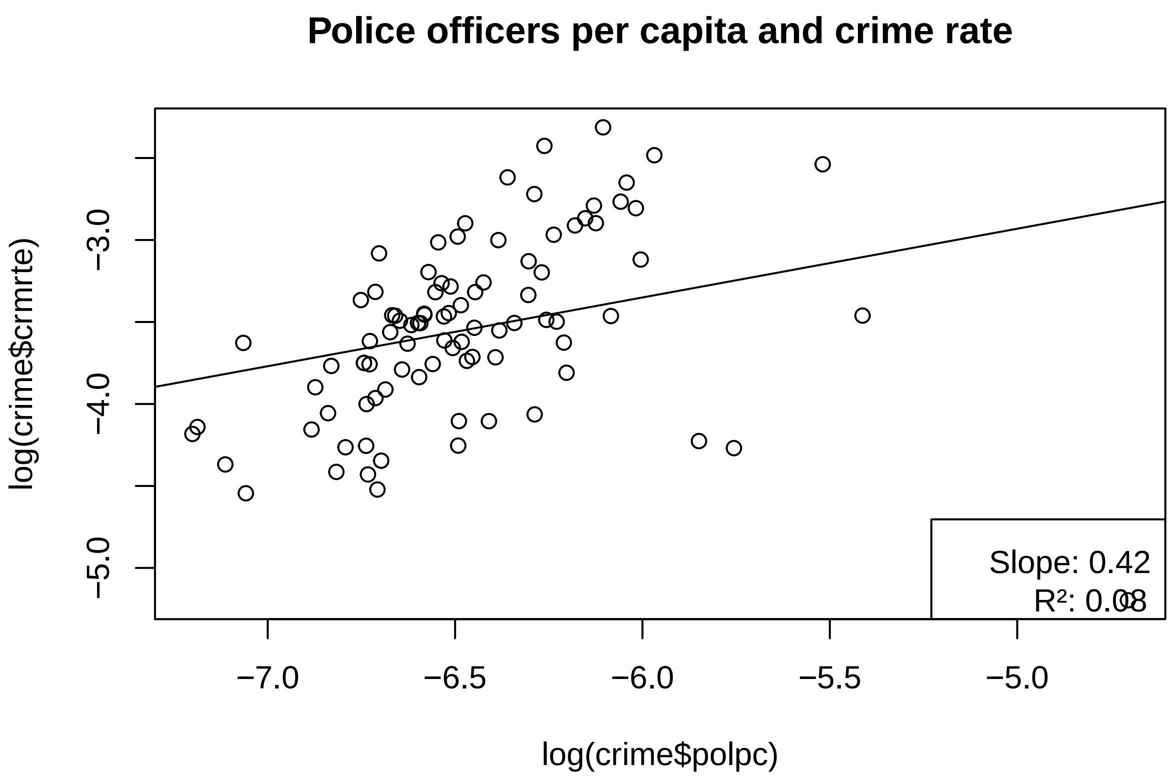
Based on the slope and R^2 statistic of the best-fit line we see that a 1% increase in probability of arrest is associated with a ~0.59% decrease in the crime rate, and that probability of arrest by itself can predict 19% of the variation in crime rate across the state.

Based on similar observations, we also apply a natural log transformation to the probability of conviction (`prbconv`) and probability of prison sentence (`prbpris`) variables.

Police officers per capita: `polpc`

A natural instinct of policy makers working to reduce crime may be to increase the number of police officers, and so it is important for them understand the existing associate between police officers per capita (`polpc`) and the crime rate. Here again we apply a log transform, as it improves the linearity of the relationship with our outcome variable.

```
model.polpc <- lm(log(crime$crmrte) ~ log(crime$polpc))
plot(log(crime$polpc), log(crime$crmrte),
     main = 'Police officers per capita and crime rate')
abline(model.polpc)
legend('bottomright', legend=r2_text(model.polpc))
```

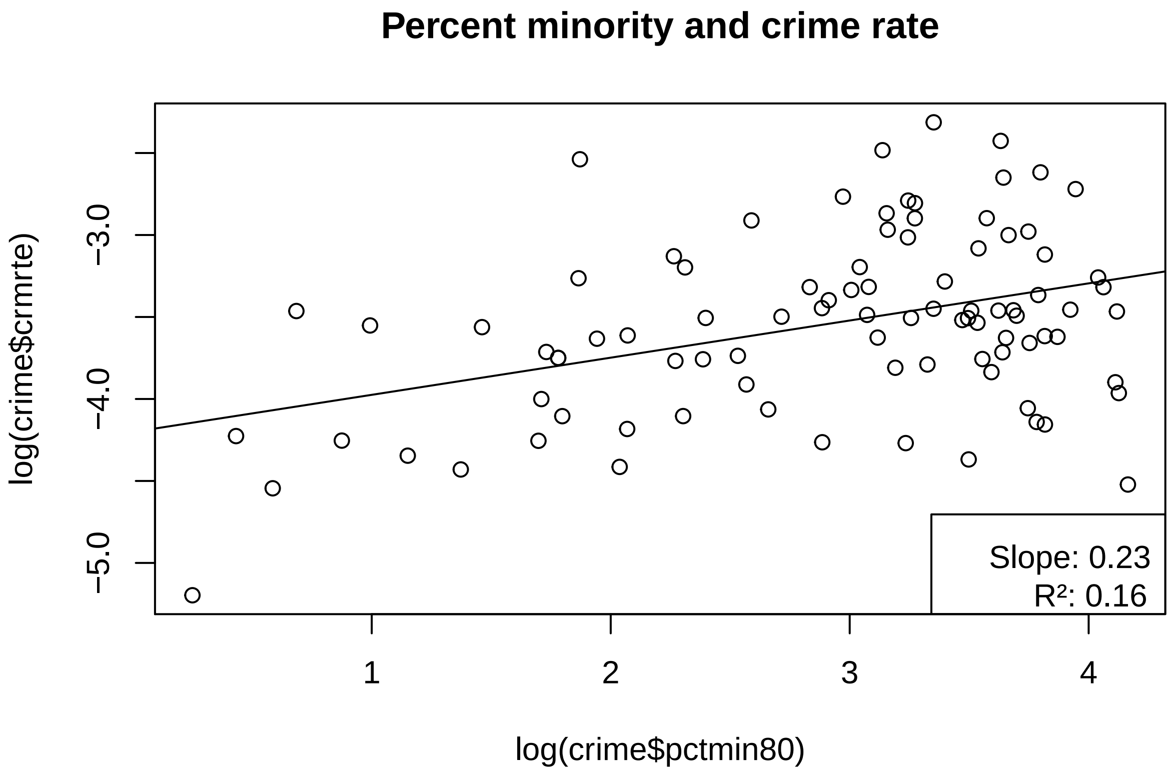


Based on the slope and R^2 statistic of the best-fit line we see that a 1% increase in police per capita is associated with a ~0.42% increase in the crime rate, and that this association predicts 8% of the variation in crime rate.

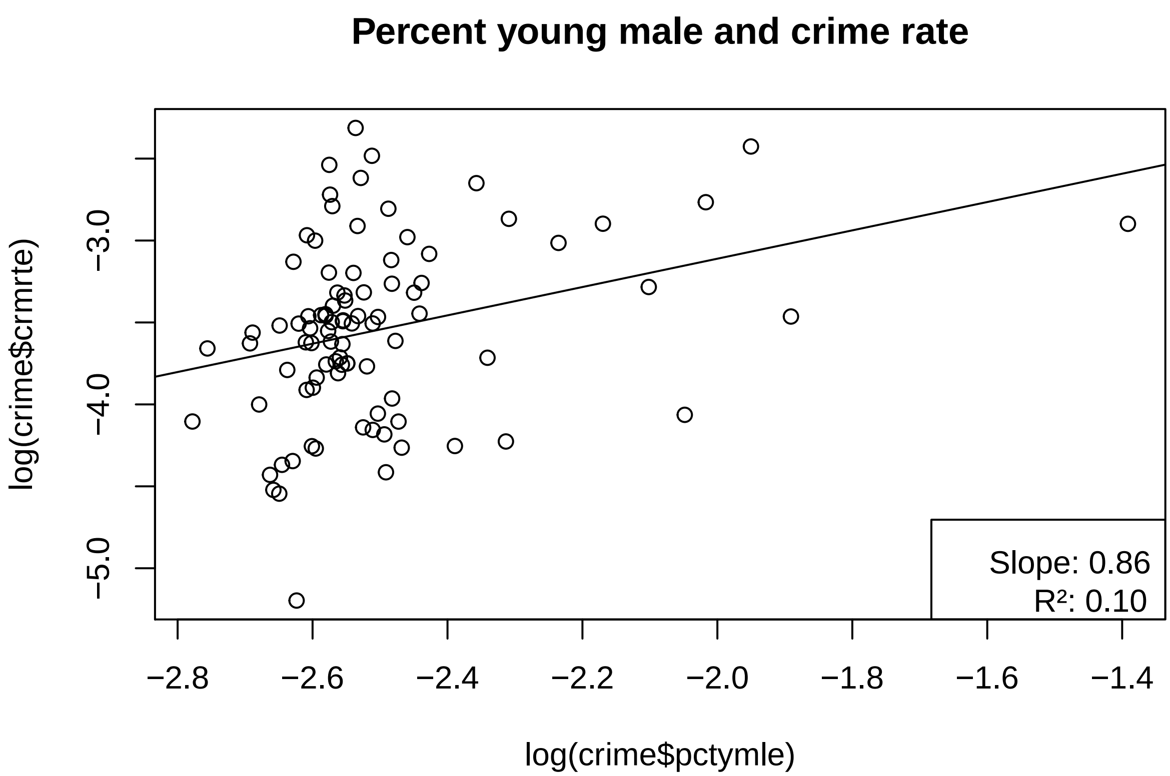
Percent young male and percent minority: `pctymle` and `pctmin80`

Furthermore, the percentages of young males and minorities in a population need to be examined as those tend to receive ongoing political focus, and the campaign should be aware of the actual relationships between those variables and crime rates. For the same reasons as discussed above, we apply a log transform to both of these variables.

```
model.pctmin80 <- lm(log(crime$crmrte) ~ log(crime$pctmin80))
plot(log(crime$pctmin80), log(crime$crmrte),
     main = 'Percent minority and crime rate')
abline(model.pctmin80)
legend('bottomright', legend=r2_text(model.pctmin80))
```




```
model.pctymle <- lm(log(crime$crmrte) ~ log(crime$pctymle))
plot(log(crime$pctymle), log(crime$crmrte),
     main = 'Percent young male and crime rate')
abline(model.pctymle)
legend('bottomright', legend=r2_text(model.pctymle))
```



Based on the slope and R^2 statistic of the best-fit line we see that a 1% increase in percentage minority is associated with a ~0.23% increase in crime rate. Similarly, a 1% increase in percentage young male is associated with a ~0.86% increase in crime rate.

For policy makers, a positive correlation between minorities and crime could be useful evidence to support policies that would lead to greater integration of cultural groups into every neighborhood. Similarly, a positive relationship between young males and crime could be used to justify funding for programs that help young males stay occupied and away from criminal elements.

Classical Linear Model Assumptions

We are now in a position to evaluate the first three classical linear model assumptions. First, the linearity observed above between the transformed key relationship variables and the transformed crime rate gives us reasonable confidence that the true relationship between these variables is approximately linear. Second, the sample under consideration is a cross section of counties in North Carolina and is assumed to be a true random sample of the population of North Carolina counties. Third, our exploratory analysis found no perfect collinearity between the variables of key interest.

Regression Models

Model 1

Our first model includes the following explanatory variables of key interest to the political campaign:

1. Natural log of probability of arrest (`prbarr`)
2. Natural log of probability of conviction given arrest (`prbconv`)
3. Natural log of police per capita (`polpc`)

These variables were selected as they are factors that a political operative can potentially influence, once they are in power.

The probability of prison (`prbpris`) and average sentence length (`avgsen`) variables are omitted purposely because they do not provide any meaningful predictive association with the crime rate. This observation is valuable for policy-makers as it could permit better decision making with respect to allocation of funds.

```
(Model1 <- lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc), data=crime))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc),
##     data = crime)
##
## Coefficients:
##  (Intercept)   log(prbarr)  log(prbconv)    log(polpc)
##      -2.4450       -0.7354       -0.4389        0.3697
```

```
sprintf('R²: %.3f', summary(Model1)$r.squared)
```

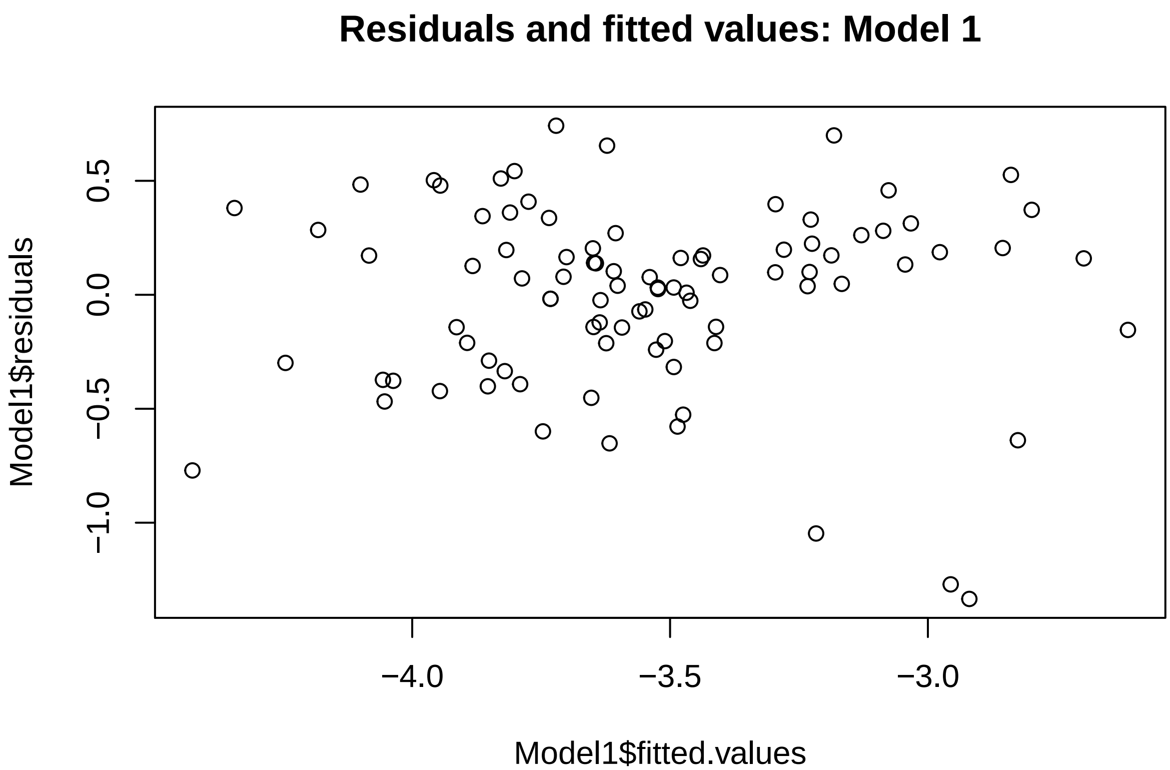
```
## [1] "R²: 0.474"
```

The model coefficients imply that, all else being equal, an increase in probability of arrest (`prbarr`) of 1% is associated with ~0.74% decrease in the crime rate. Additionally, all else being equal, an increase in the probability of conviction given arrest (`prbconv`) of 1% is associated with a ~0.43% decrease in the crime rate. Surprisingly, it is also found that a 1% increase in police officers per capita is associated with a ~0.36% *increase* in crime rate. The model's R^2 value indicates that 47.4% of the variation in crime rate can be explained by these three factors alone.

Classical Linear Model Assumptions

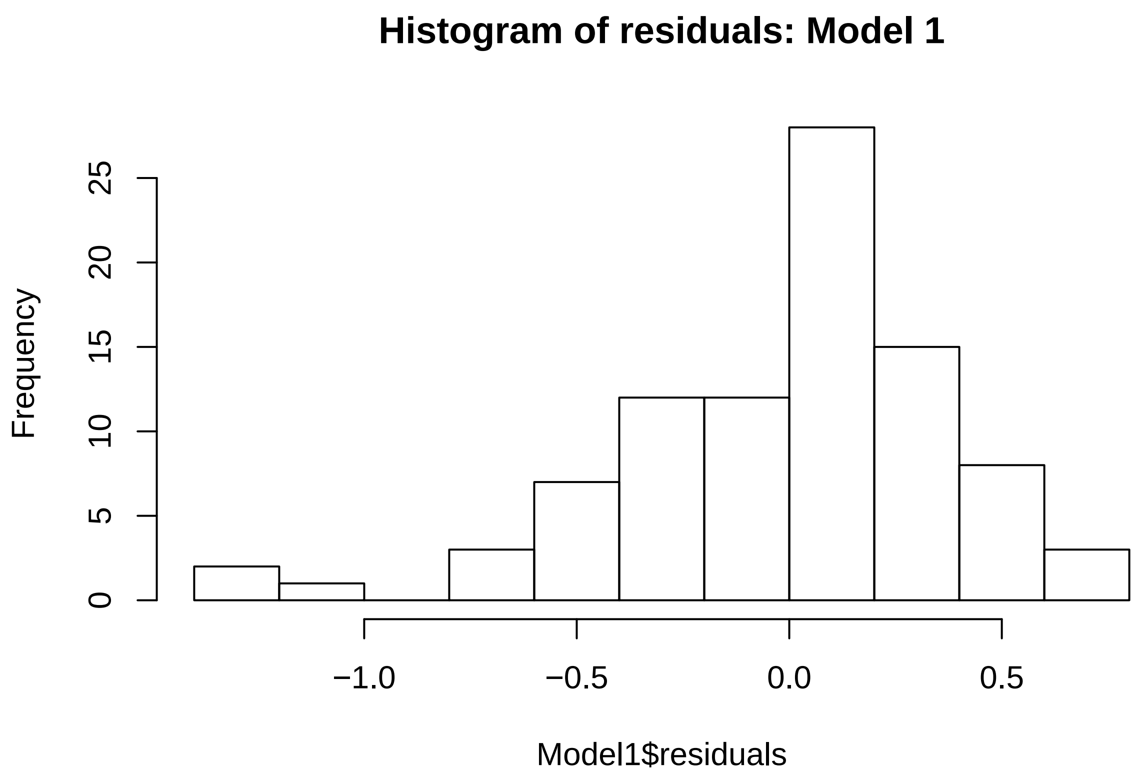
The residual vs. fitted value plot shows the residuals are centered on 0 (exogeneity) with relatively constant variance (homoskedasticity) across the range of fitted values.

```
plot(Model1$fitted.values, Model1$residuals,
     main = 'Residuals and fitted values: Model 1')
```



This histogram of residuals shows that the residuals are distributed approximately normally with mean 0 (with the exception of a few outlying points with larger negative residuals).

```
hist(Model1$residuals, main = 'Histogram of residuals: Model 1')
```



Model 2

The second model includes the following terms:

1. Natural log of probability of arrest (`prbarr`)
2. Natural log of probability of conviction given arrest (`prbconv`)
3. Natural log of police per capita (`polpc`)
4. Natural log of percent minority in 1980 (`pctmin80`)
5. Natural log of population density (`density`)

The log of density term was added to Model 2 to improve accuracy because the exploratory data analysis revealed that there is a positive correlation between density and crime rate. This correlation has been noted by other researchers (Geoffrey West in "Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life in Organisms, Cities, Economies, and Companies"). This is a useful variable as policy makers could encourage development and housing policies that lead to lower population density such as improving transit options to suburban areas.

The log of the percent minority population was added to improve accuracy because, again, exploratory data analysis revealed that there is a positive correlation between and crime rate and this factor.

```r
(Model2 = lm(log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc) +
               log(pctmin80) + log(density), data = crime))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc) +
##     log(pctmin80) + log(density), data = crime)
##
## Coefficients:
##   (Intercept)    log(prbarr)   log(prbconv)     log(polpc)  log(pctmin80)
##       -3.0217        -0.4403        -0.3506         0.3218         0.2479
```

```
## log(density)
##       0.2794
```

```r
sprintf('R²: %.3f', summary(Model2)$r.squared)
```

```
## [1] "R²: 0.794"
```

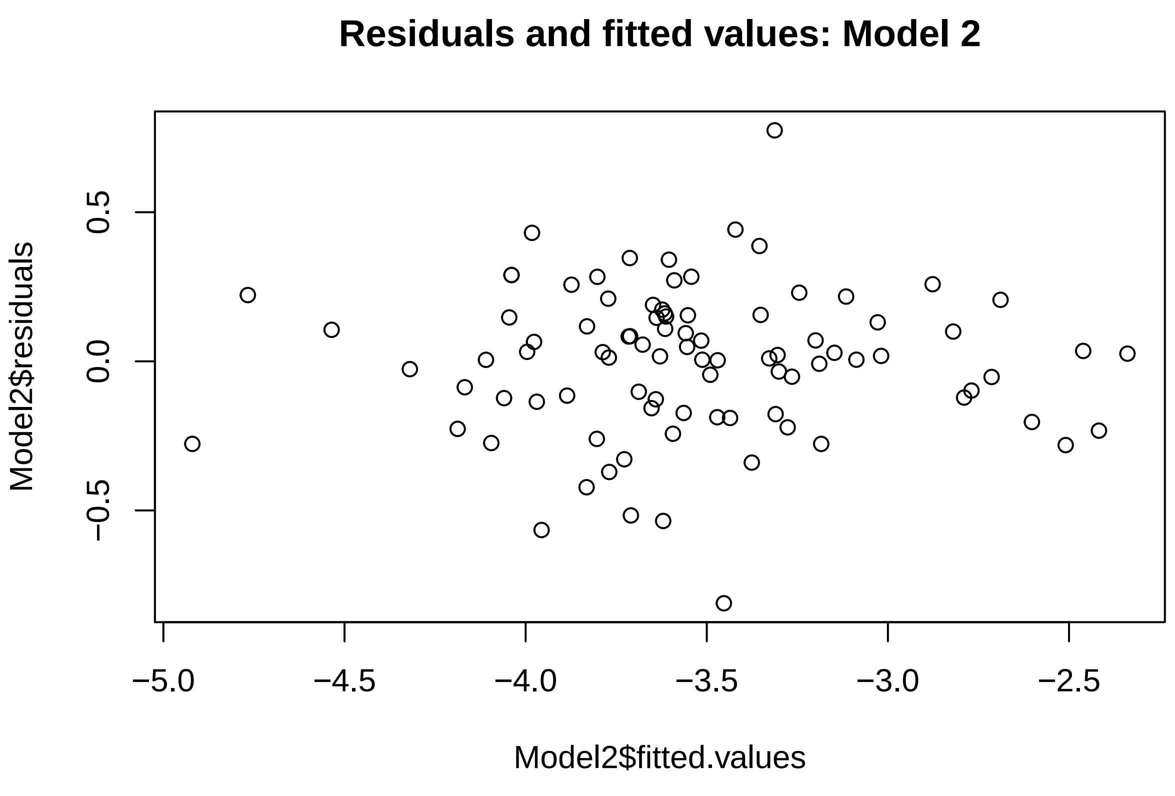
The `prbarr`, `prbconv`, and `polpc` variables retain their same direction of associate as in Model 1, although the magnitude of the associations are reduced somewhat. In addition, we see that a 1% increase in the percent minority population is associated with a ~0.25% increase in the crime rate, and that a 1% increase in the population density is associated with a ~0.28% increase in crime rate.

The model's R^2 value indicates that 79.4% of the variation in crime rate can be explained by these 5 factors.

Classical Linear Model Assumptions

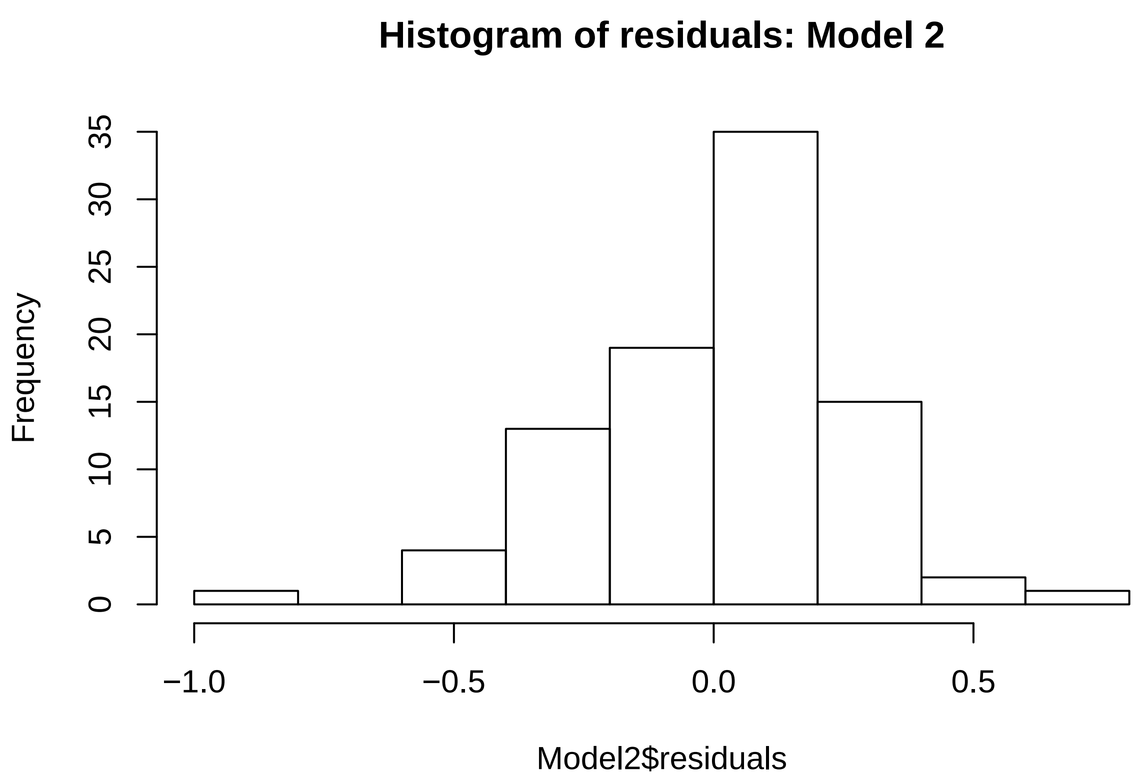
The residual vs. fitted value plot shows the residuals are centered on 0 (exogeneity) with relatively constant variance (homoskedasticity) across the range of fitted values.

```r
plot(Model2$fitted.values, Model2$residuals,
     main = 'Residuals and fitted values: Model 2')
```



This histogram of residuals shows that the residuals are distributed normally with mean 0.

```r
hist(Model2$residuals, main = 'Histogram of residuals: Model 2')
```



Model 3

Model 3 was constructed using every variable with the exception of the “west”, “central”, and “urban” variables. We believed that these variables are already correlated with density and they do not have any intrinsic value in informing policy.

```
(Model3 = lm(log(crmrte) ~  log(prbarr) + log(prbconv) + log(polpc) +
               log(pctmin80) + log(density) + log(prbpris) +
               log(pctymle) + avgsen + wser + county +  taxpc + wcon +
               wtuc + wtrd + wfir + wmfg + wfed + wsta + wloc + mix,
             data = crime))
```

```
##
## Call:
## lm(formula = log(crmrte) ~ log(prbarr) + log(prbconv) + log(polpc) +
##     log(pctmin80) + log(density) + log(prbpris) + log(pctymle) +
##     avgsen + wser + county + taxpc + wcon + wtuc + wtrd + wfir +
##     wmfg + wfed + wsta + wloc + mix, data = crime)
##
## Coefficients:
##   (Intercept)    log(prbarr)   log(prbconv)     log(polpc)  log(pctmin80)
##    -2.1903211     -0.4104629     -0.2546735      0.3600833      0.2218646
##  log(density)   log(prbpris)   log(pctymle)         avgsen           wser
##     0.2627645     -0.1782582      0.2041088     -0.0296140     -0.0020262
##        county          taxpc           wcon           wtuc           wtrd
##     0.0002577      0.0051699      0.0001841      0.0002615      0.0008327
##          wfir           wmfg           wfed           wsta           wloc
##    -0.0010893     -0.0001383      0.0016809     -0.0003886      0.0000034
##           mix
```

```
## -0.0522225
```

```r
sprintf('R²: %.3f', summary(Model3)$r.squared)
```

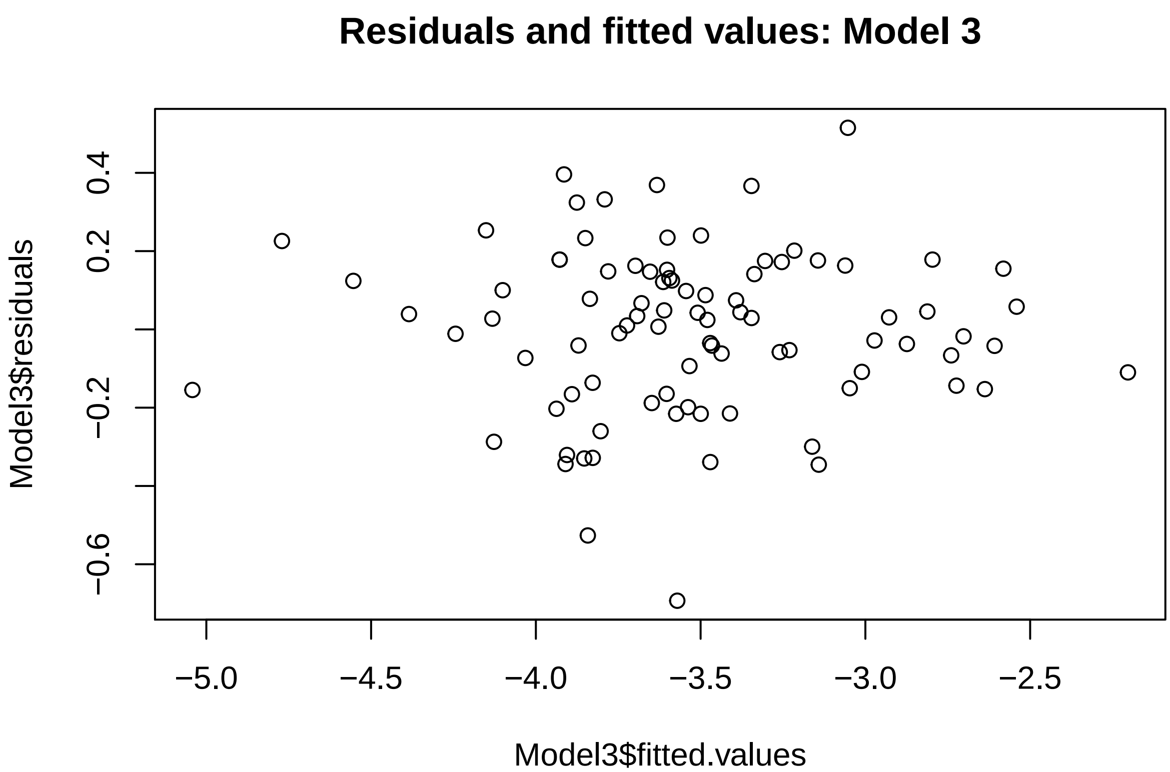
```
## [1] "R²: 0.848"
```

By including all the remaining variables, the r-squared value of Model 3 improved by ~5% percentage points over Model 2. This compares with a ~32% percentage point improvement between Model 2 and Model 1. The analysts at Berkeley Analytica conclude that the improved fit of Model 3 is not justified by the additional data. In short, Model 2 is more valuable as Model 3 may be approaching over-fitting.

CLM Assumptions

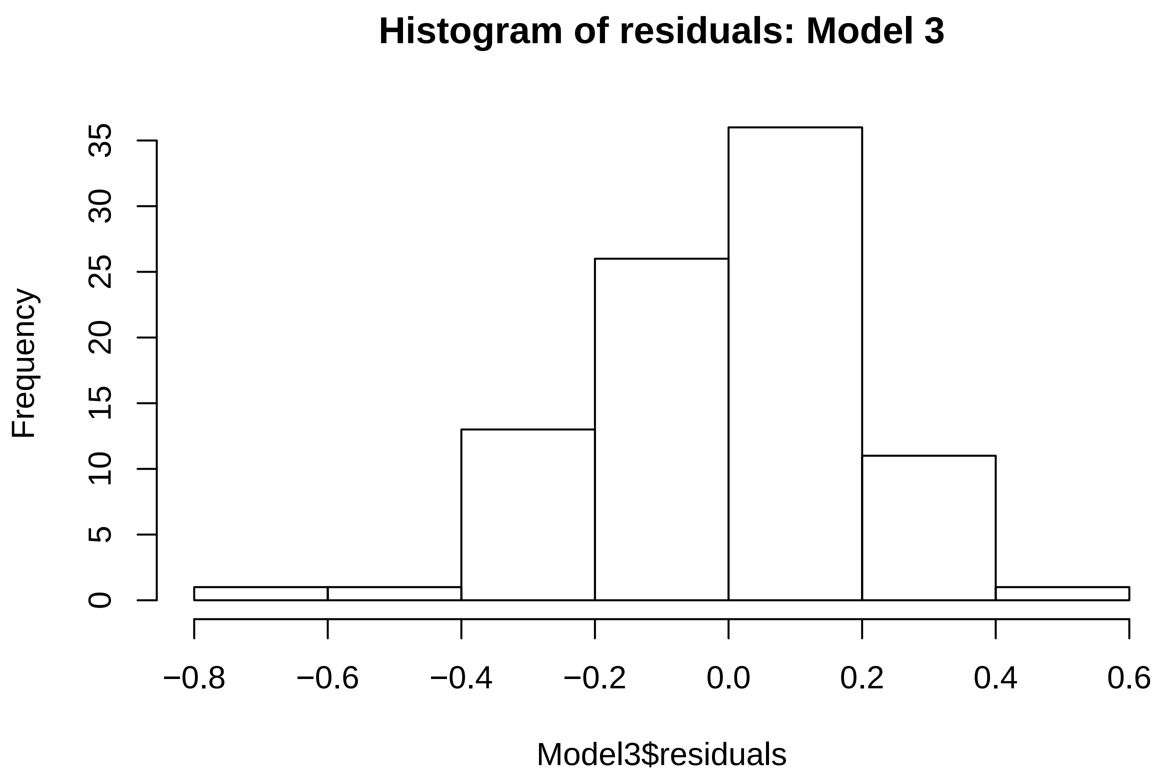
The residual vs. fitted value plot shows the residuals are centered on 0 (exogeneity) with relatively constant variance (homoskedasticity) across the range of fitted values.

```r
plot(Model3$fitted.values, Model3$residuals, main = 'Residuals and fitted values: Model 3')
```



This histogram of residuals shows that the residuals are distributed approximately normally with mean 0, although there is a slight negative skew.

```r
hist(Model3$residuals, main = 'Histogram of residuals: Model 3')
```



Models Table

```
stargazer(Model1, Model2, Model3, type = "latex",
          report = "vc", # Don't report errors, since we haven't covered them
          title = "Linear Models Predicting Crime Rate",
          keep.stat = c("rsq", "n"),
          omit.table.layout = "n",
          header=FALSE) # Omit more output related to errors
```

Omitted Variables Discussion

The results from the polpc (Police per capita) variable is counter-intuitive at first glance. However, on further examination, it becomes clear that both crime rate and police per capita increase with increasing density, however, crime rate increases faster with density than police per capita. The net result is that there is a positive correlation between police and crime. This does not, however, indicate that increasing police on the job will increase the crime rate.

Conclusion

Based on analysis of the data in this study, the political campaign could argue that to decrease crime rates, policing should be geared towards more and faster arrests. Similarly, the criminal justice system should be geared towards more and faster convictions with shorter prison sentences. Shortening prison sentences would allow for funds to be freed for use in other areas. The deterrent seems to be in arrests and conviction, not as much in being sent to prison, or based on how long a prison sentence is.

Similarly, the political campaign could push for funding programs that keep young males busy and away from criminal elements.

Table 1: Linear Models Predicting Crime Rate

	Dependent variable:		
	log(crmrte)		
	(1)	(2)	(3)
log(prbarr)	−0.735	−0.440	−0.410
log(prbconv)	−0.439	−0.351	−0.255
log(polpc)	0.370	0.322	0.360
log(pctmin80)		0.248	0.222
log(density)		0.279	0.263
log(prbpris)			−0.178
log(pctymle)			0.204
avgsen			−0.030
wser			−0.002
county			0.0003
taxpc			0.005
wcon			0.0002
wtuc			0.0003
wtrd			0.001
wfir			−0.001
wmfg			−0.0001
wfed			0.002
wsta			−0.0004
wloc			0.00000
mix			−0.052
Constant	−2.445	−3.022	−2.190
Observations	91	90	89
R^2	0.474	0.794	0.848

Further studies are needed to understand if value is created for society by having a lower population density and if this value could be realized by improving transportation options to allow for more suburban developments in favor of urban developments.

Finally, further studies could be performed to see if improving the mix of minorities in neighborhoods helps to reduce crime rates.