W203-2, Week 15, Lab 4

Tako Hisada

12/17/2017

```
library(car)
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

Introduction

The United States is known to have the highest prison population in the world. Our team has been hired by a political campaign to provide research in identifying factors that influence the probability of getting sentenced (*probsen*) for the offences committed. By identifying these factors, the team hopes to help the campaign formulate possible legislative actions that the government could undertake in reducing such crimes and hence the number of inmates in the prisons.

Initial exploratory analysis

The file crime.csv contains crime statistics for a selection of counties. While it is possible that there are factors not included in the dataset that are contributing to jail sentences, we have a pretty comprehensive set of variables given in the dataset ranging from crime, geography, economic, and demographics of the counties included in the dataset each of which we will delve into shortly.

```
Data <- read.csv('crime_v2_updated.csv')
head(Data)
```

```
##   X county year     crime  probarr  probsen  probconv avgsen     police
## 1 1      1   88 0.0356036 0.436170 0.298270 0.5275960   6.71 0.00182786
## 2 2      3   88 0.0152532 0.450000 0.132029 1.4814800   6.35 0.00074588
## 3 3      5   88 0.0129603 0.600000 0.444444 0.2678570   6.76 0.00123431
## 4 4      7   88 0.0267532 0.435484 0.364760 0.5254240   7.14 0.00152994
## 5 5      9   88 0.0106232 0.442623 0.518219 0.4765630   8.22 0.00086018
## 6 6     11   88 0.0146067 0.500000 0.524664 0.0683761  13.00 0.00288203
##     density      tax west central urban   pctmin   wagecon   wagetuc
```

```
## 1 2.4226327 30.99368    1       0     0 20.21870 281.4259 408.7245
## 2 1.0463320 26.89208    1       0     0  7.91632 255.1020 376.2542
## 3 0.4127659 34.81605    0       1     0  3.16053 226.9470 372.2084
## 4 0.4915572 42.94759    1       0     0 47.91610 375.2345 397.6901
## 5 0.5469484 28.05474    0       1     0  1.79619 292.3077 377.3126
## 6 0.6113361 35.22974    0       1     0  1.54070 250.4006 401.3378
##    wagetrd  wagefir  wageser wagemfg wagefed wagesta wageloc        mix
## 1 221.2701 453.1722 274.1775  334.54  477.58  292.09  311.91 0.08016878
## 2 196.0101 258.5650 192.3077  300.38  409.83  362.96  301.47 0.03022670
## 3 229.3209 305.9441 209.6972  237.65  358.98  331.53  281.37 0.46511629
## 4 191.1720 281.0651 256.7214  281.80  412.15  328.27  299.03 0.27362204
## 5 206.8215 289.3125 215.1933  290.89  377.35  367.23  342.82 0.06008584
## 6 187.8255 258.5650 237.1507  258.60  391.48  325.71  275.22 0.31952664
##        ymale
## 1 0.07787097
## 2 0.08260694
## 3 0.07211538
## 4 0.07353726
## 5 0.07069755
## 6 0.09891920
```

```r
n <- nrow(Data)
num_cols <- ncol(Data)
```

head() confirms that the data has been succesfully loaded. The dataset contains 26 columns (variables) and 90 rows. This is sufficiently large enough to assume CLT.

```r
# Check for NAs
for(i in names(Data)){
  val <- Data[[i]][is.na(Data[[i]])]
  if(length(val)) {
    sprintf("%s: %d NA row(s) found", i, length(val))
  }
}
```

No NAs are found in the dataset given.

Individual variable analysis

X

This is just an index variable and hence no analysis is required.

Country identifier

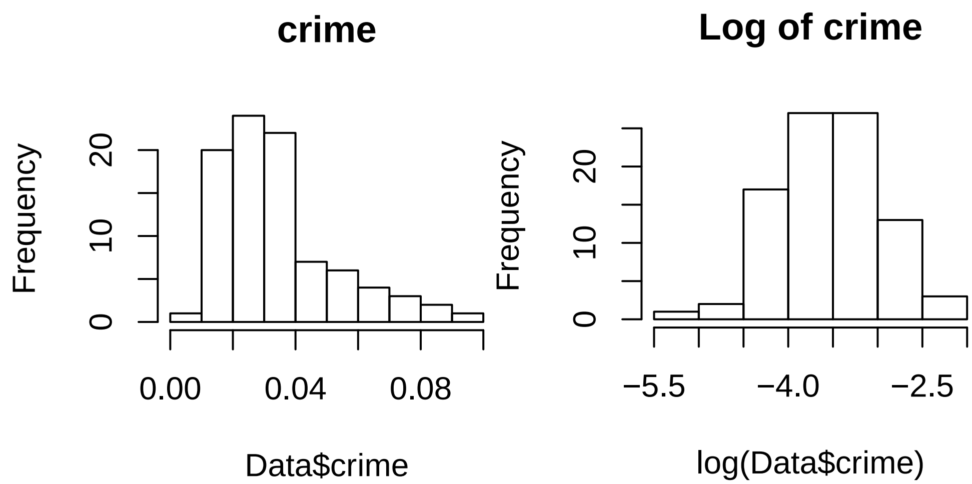
This is just an identifier and hence no analysis is required.

Year

This is just the year when this data was collected and it is simply 88 for all rows. No analysis requried.

Crime committed per person

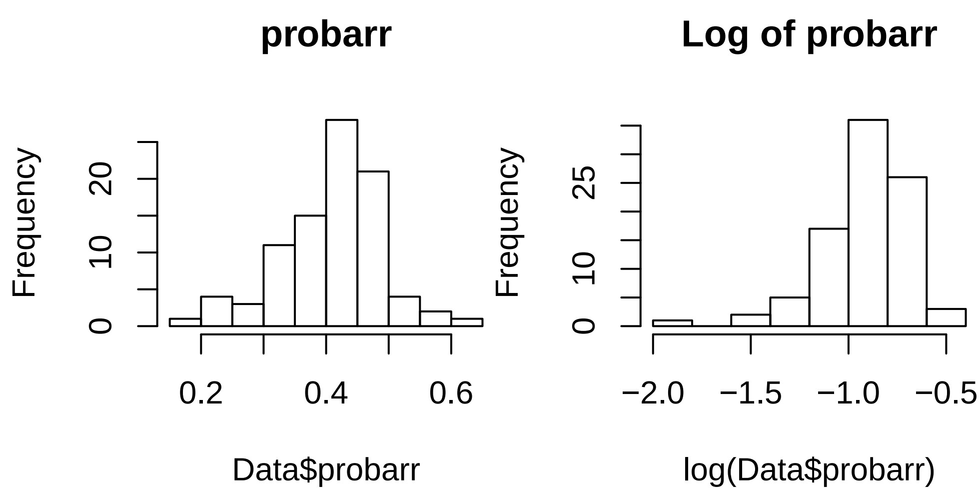
```r
hist(Data$crime, main = "crime")
hist(log(Data$crime), main = "Log of crime")
```



The histogram is positively skewed. No extreme outliers observed. The histogram becomes more normal when log() is applied.

'Probability' of arrest

```r
hist(Data$probarr, main = "probarr")
hist(log(Data$probarr), main = "Log of probarr")
```

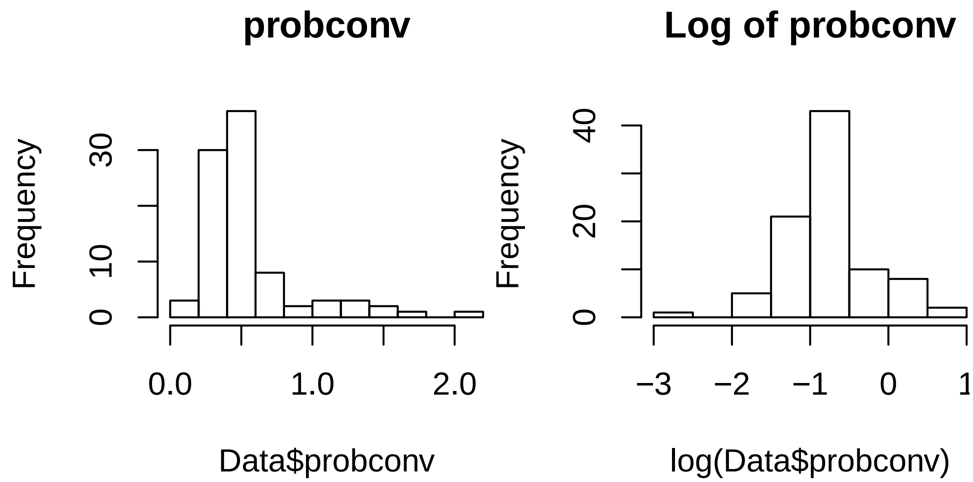


The histogram is relatively normal. No extreme outliers observed. The histogram actually becomes less normal when log() is applied.

'Probability' of conviction

```r
hist(Data$probconv, main = "probconv")
hist(log(Data$probconv), main = "Log of probconv")
(length(Data$probconv[Data$probconv > 1]))
```

```
## [1] 10
```

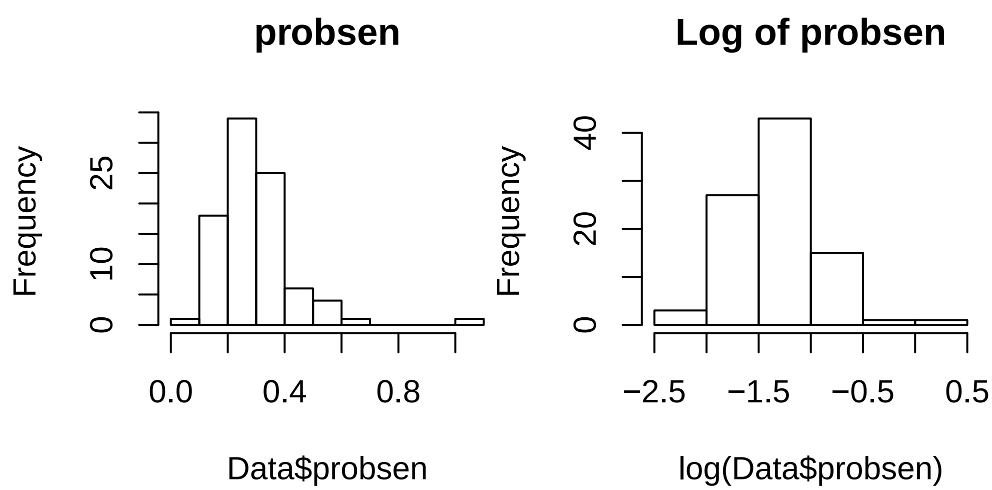


The histogram is positively skewed with extreme outliers (10 items over 1). The histogram becomes more normal when log() is applied.

'Probability' of prison sentence

```r
hist(Data$probsen, main = "probsen")
hist(log(Data$probsen), main = "Log of probsen")
(length(Data$probsen[Data$probsen > 1]))
```

```
## [1] 1
```



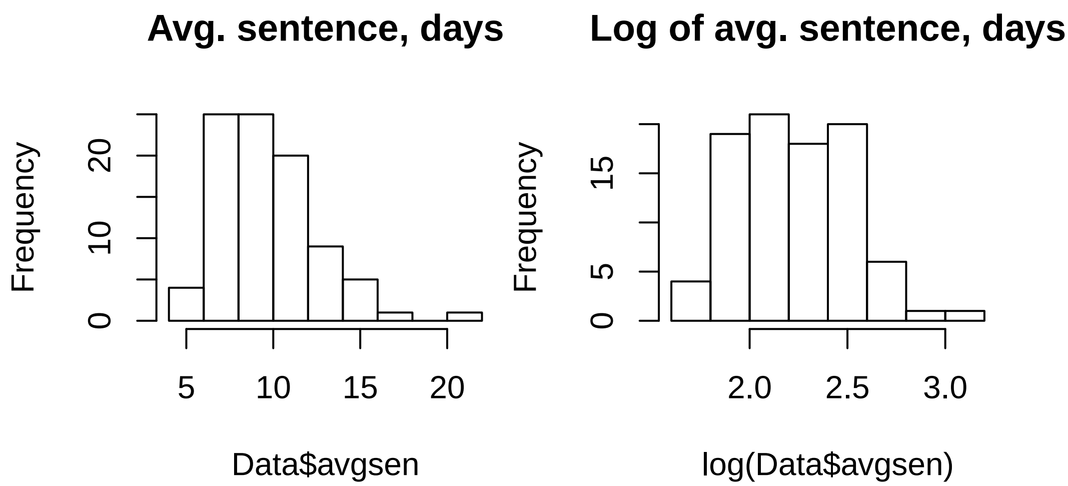
The histogram is relatively normal with an exception of one extreme outlier (10 items over 1). The histogram becomes more normal when log() is applied.

Avg. sentence, days

```r
hist(Data$avgsen, main = "Avg. sentence, days")
(length(Data$probsen[Data$avgsen > 20]))
```

```
## [1] 1
```

```
hist(log(Data$avgsen), main = "Log of avg. sentence, days")
```



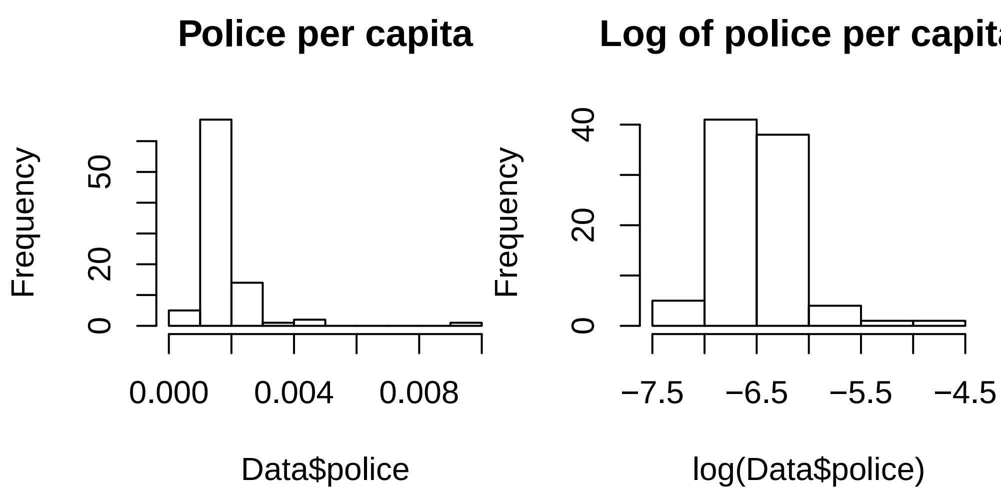
The histogram is slightly positively skewed with 1 outlier (20 >). The histogram becomes more normal when log() is applied.

Police per capita

```
hist(Data$police, main = "Police per capita")
(length(Data$probsen[Data$police > 0.009]))
```

```
## [1] 1
```

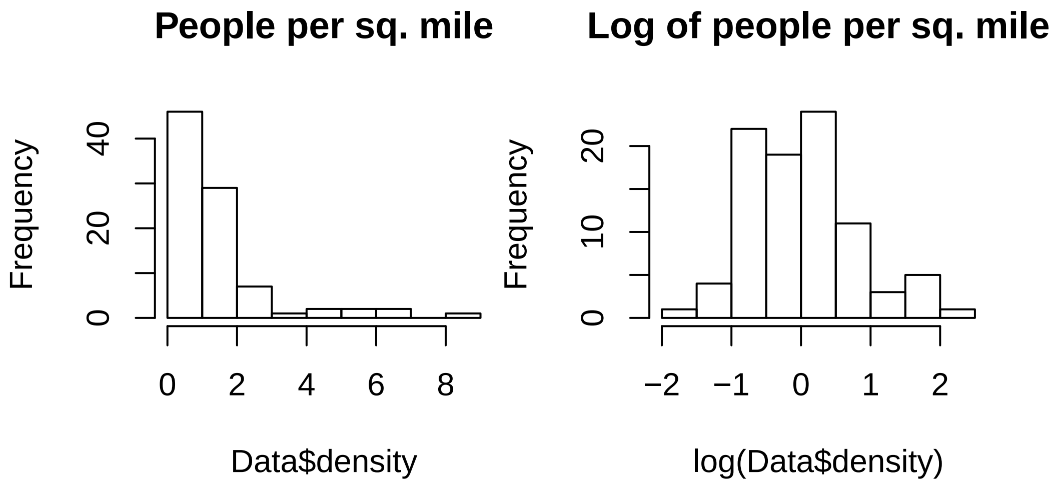
```
hist(log(Data$police), main = "Log of police per capita")
```



The histogram is positively skewed with 1 outlier. The histogram becomes slightly more normal when log() is applied.

People per sq. mile

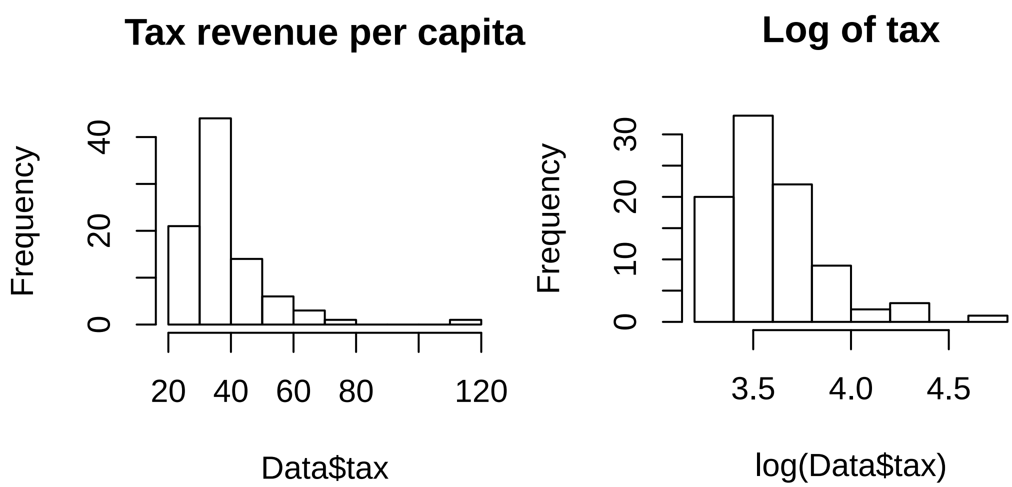
```
hist(Data$density, main = "People per sq. mile")
hist(log(Data$density), main = "Log of people per sq. mile")
```



The histogram is positively skewed. The histogram becomes more normal when log() is applied.

Tax revenue per capita

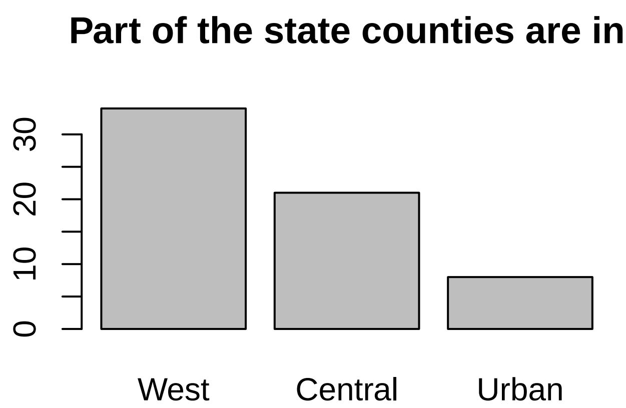
```r
hist(Data$tax, main = "Tax revenue per capita")
hist(log(Data$tax), main = "Log of tax")
```



The histogram is positively skewed. The histogram becomes slightly more normal when log() is applied however is still positively skewed.

West/Central/Urban

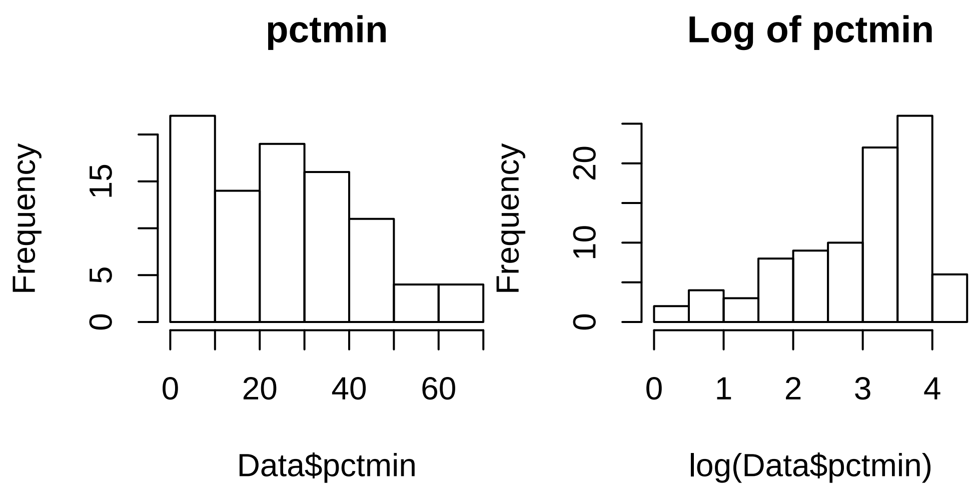
```r
barplot(c(sum(Data$west), sum(Data$central), sum(Data$urban)),
        names.arg = c("West", "Central", "Urban"), main = "Part of the state counties are in")
sum_geo <- sum(Data$west) + sum(Data$central) + sum(Data$urban)
```



Dummy variables indicating whether or not a given county is in the western/central/urban part of the state. Interestingly, the sum of the 3 regions only add up to 63 which is considerably less than our n of 90. There are many 27 counties in the dataset do not fall under any of these regions.

Proportion that is minority or nonwhite

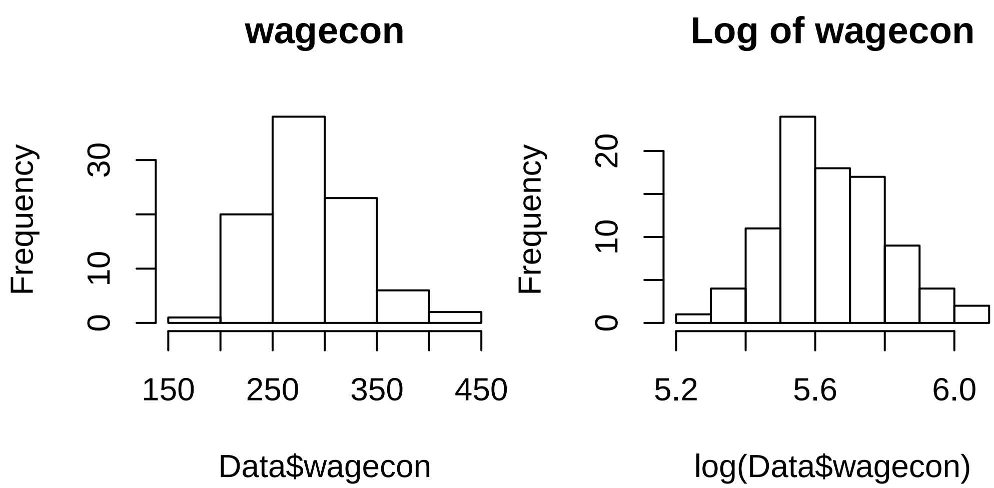
```r
hist(Data$pctmin, main = "pctmin")
hist(log(Data$pctmin), main = "Log of pctmin")
```



The histogram is positively skewed. The histogram becomes more normal when log() is applied althogh it is still negatively skewed.

Weekly wage, construction

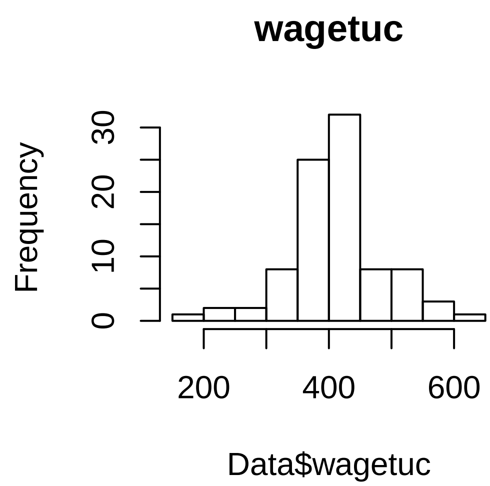
```r
hist(Data$wagecon, main = "wagecon")
hist(log(Data$wagecon), main = "Log of wagecon")
```



The histogram is pretty normal. The histogram becomes more normal when log() is applied.

Weekly wage, transportation, utilities, communications

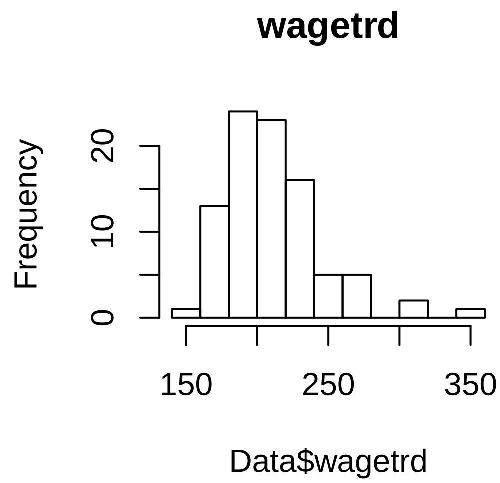
```r
hist(Data$wagetuc, main = "wagetuc")
```



The histogram is relatively normal.

Weekly wage, wholesale, retail trade

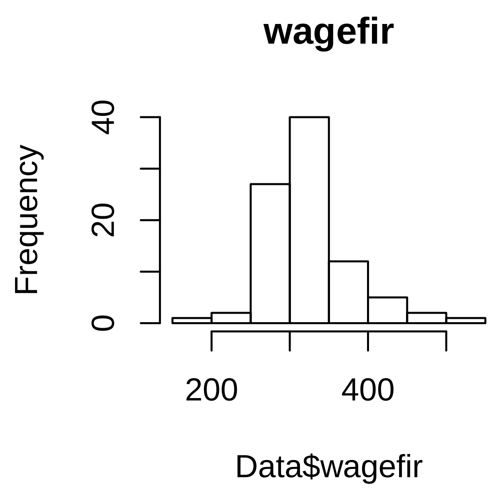
```r
hist(Data$wagetrd, main = "wagetrd")
```



The histogram is relatively normal with some outliers.

Weekly wage, finance, insurance and real estate

```r
hist(Data$wagefir, main = "wagefir")
```

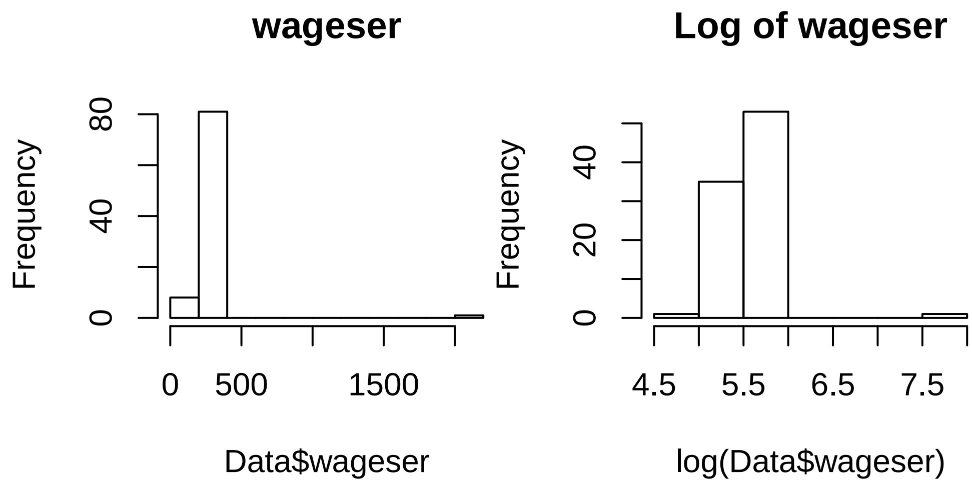


The histogram is relatively normal.

Weekly wage, service industry

```r
hist(Data$wageser, main = "wageser")
hist(log(Data$wageser), main = "Log of wageser")
max(Data$wageser)
```

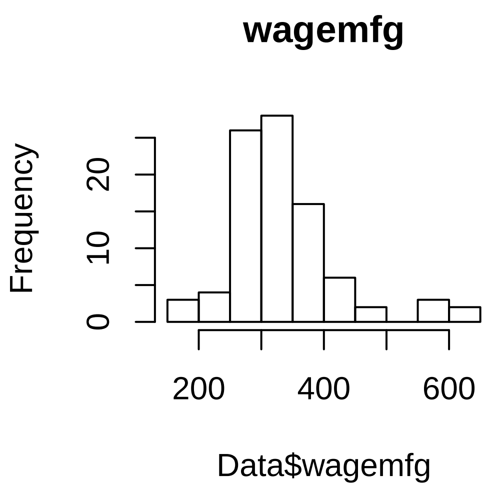
```
## [1] 2177.068
```



The histogram is positively skewed with one extreme outlier. The histogram becomes slightly more normal when log() is applied.

Weekly wage, manufacturing

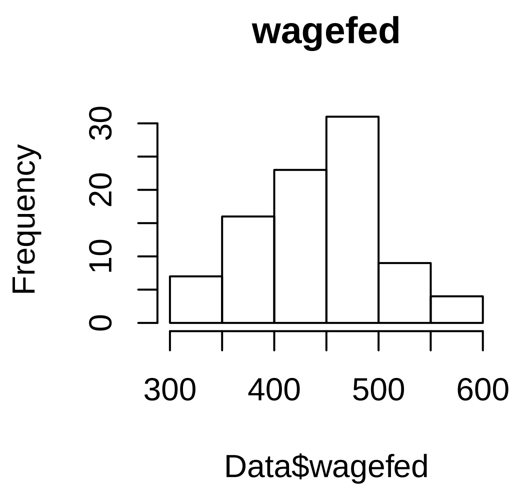
```r
hist(Data$wagemfg, main = "wagemfg")
```



The histogram is relatively normal but with some outiers.

Weekly wage, federal employees

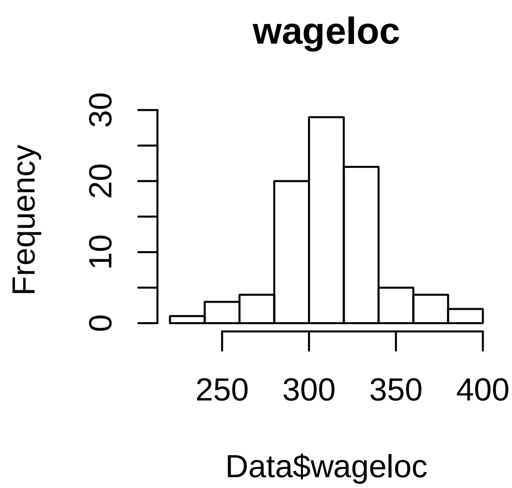
```r
hist(Data$wagefed, main = "wagefed")
```



The histogram is relatively normal.

Weekly wage, local government employees

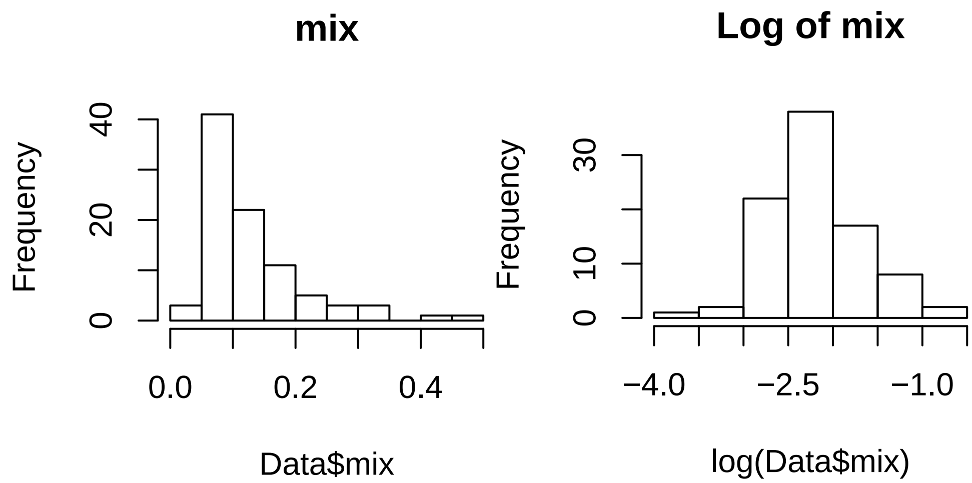
```
hist(Data$wageloc, main = "wageloc")
```



The histogram is pretty normal.

Ratio of face to face/all other crimes

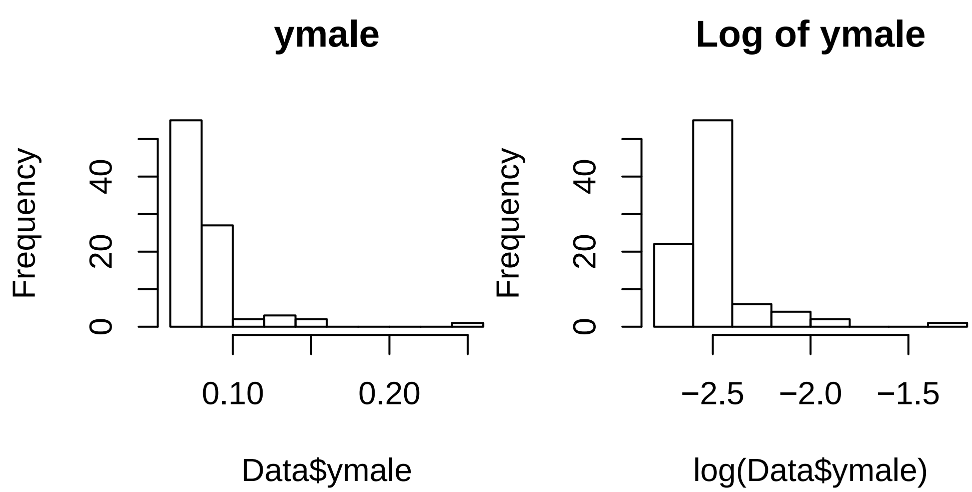
```
hist(Data$mix, main = "mix")
hist(log(Data$mix), main = "Log of mix")
```



The histogram is positively skewed. The histogram becomes more normal when log() is applied.

Proportion of county males between the ages of 15 and 24

```
hist(Data$ymale, main = "ymale")
hist(log(Data$ymale), main = "Log of ymale")
```

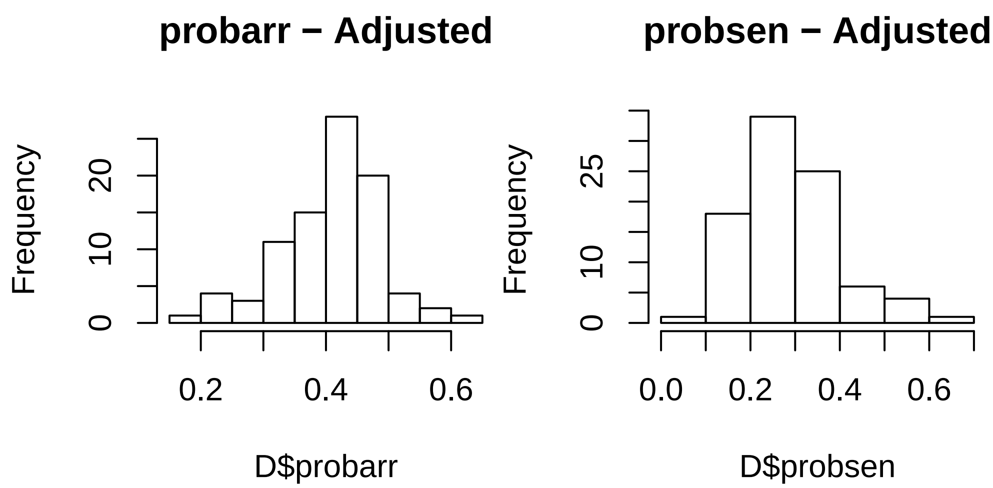


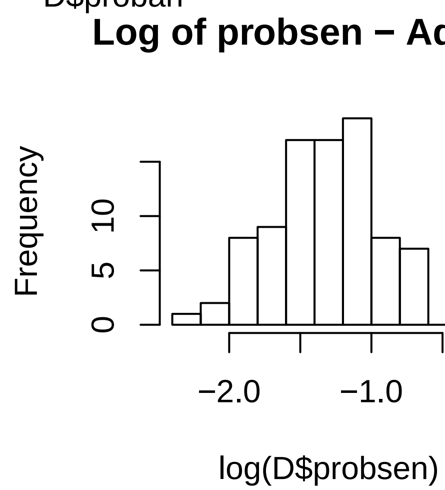
The histogram is positively skewed. The histogram becomes more normal when log() is applied however it is still positively skewed.

Variable transformations

probarr and *probsen* contain values > 1 which are difficult to intrepret. We are omitting them from our analysis.

```
D <- Data[Data$probarr < 1 & Data$probsen < 1,]
hist(D$probarr, main = "probarr - Adjusted")
hist(D$probsen, main = "probsen - Adjusted")
hist(log(D$probsen), main = "Log of probsen - Adjusted")
```





Without the outliers, the 2 histograms are looking relatively normal. *probsen* is still looking positively skewed so we are taking the log of it which makes the distribution more normal.

Now we will apply log() to the below variables as they are not very normally distributed and store them in the newly created dataframe D so they will be available for later analysis.

```
D$logcrime <- log(D$crime)
D$logavgsen <- log(D$avgsen)
D$logpolice <- log(D$police)
D$logprobconv <- log(D$probconv)
D$logprobsen <- log(D$probsen)
D$logdensity <- log(D$density)
D$logtax <- log(D$tax)
D$logpctmin <- log(D$pctmin)
D$logwagecon <- log(D$wagecon)
D$logwageser <- log(D$wageser)
D$logmix <- log(D$mix)
D$logymale <- log(D$ymale)
```

Models

The team wants to explore how much is accounted for by the crime and police-related variables such as number of crimes committed (*crime*), police per capita (*police*) and ratio of face-to-face/all other crimes (*mix*) and how much of it can be attributed to other demographic variables such as race, gender, age, economic standings (wages).

Proposed Model 1 - Minimum specification

Crime-related variables: We can intuitively anticipate *probsen* to go up as *crime*, *police* and *mix* increase. My intuition would be *probsen* and *avgsen* to have a positive correlation as increased *avgsen* would suggest there would be more severe crimes happening in a given county.

Probability variables: I expect the other 2 probability variables *probarr* and *probconv* to have strong correlations with probsen and hence they will also be included in the model so we can measure how much influence the other variables have on *probsen* holding *probarr* and *probconv* fixed.

For this initial model, we will exclude the other demographic variables.

$$log(probsen) = \beta_0 + \beta_1 log(crime) + \beta_2 probarr + \beta_3 log(probconv) + \beta_4 log(avgsen)$$
$$+\beta_5 log(police) + \beta_6 log(mix) + u$$

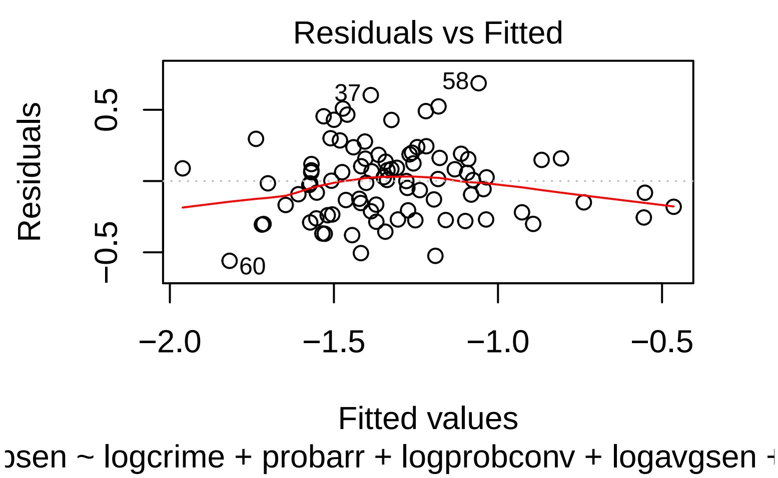
```
model1 <- lm(logprobsen ~ logcrime + probarr + logprobconv + logavgsen + logpolice + logmix, data = D)
```

CLM Assessment

CLM 1 - A linear model

The model is specified such that the dependent variable is a linear function of the explanatory variables.

```
plot(model1, which=1)
```



Ther is no non-linear relationship observed in the Residuals vs Fitted plot.

Is the assumption valid? **Yes**

CLM 2 - Random samling

As the dataset has been provided for a selection of counties, the data is not truly randomly sampled. We are not given much information about how the data in the CSV file has been collected. We will assume here that the data has been collected from the relevant random samples in these counties.

Is the assumption valid? **Yes**

CLM 3 - Multicollinearity

```
X <- data.matrix(subset(
  D, select = c("logprobsen", "logcrime", "probarr", "logprobconv", "logavgsen", "logpolice", "logmix"))
(Cor = cor(X))
```

```
##              logprobsen     logcrime     probarr logprobconv   logavgsen
## logprobsen   1.00000000 -0.360492812 -0.04064202 -0.31311633 -0.12188311
## logcrime    -0.36049281  1.000000000  0.06321588 -0.32628681  0.13418145
## probarr     -0.04064202  0.063215878  1.00000000 -0.02533560 -0.17225398
## logprobconv -0.31311633 -0.326286811 -0.02533560  1.00000000 -0.05986352
## logavgsen   -0.12188311  0.134181452 -0.17225398 -0.05986352  1.00000000
## logpolice   -0.16102212  0.542713183 -0.05647614 -0.29508551  0.29487049
## logmix       0.56189540 -0.006115974  0.09256607 -0.38424287 -0.13105421
##               logpolice       logmix
## logprobsen  -0.16102212  0.561895402
## logcrime     0.54271318 -0.006115974
## probarr     -0.05647614  0.092566073
## logprobconv -0.29508551 -0.384242869
## logavgsen    0.29487049 -0.131054206
## logpolice    1.00000000  0.062667253
## logmix       0.06266725  1.000000000
```

We are not seeing any obvious signs of multicollinearity. We will now compute VIF.

```
vif(model1)
```

```
##    logcrime     probarr logprobconv   logavgsen   logpolice      logmix
##    1.534288    1.049640    1.368897    1.151492    1.574976    1.241359
```

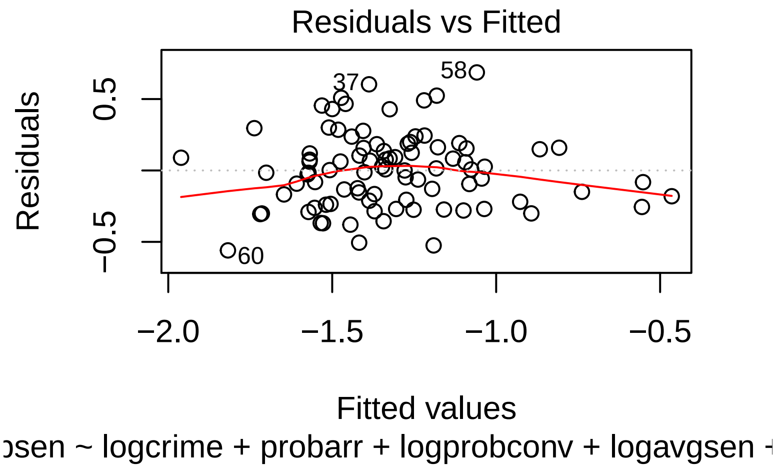
The VIF is < 4 and R is not flagging perfect multicollinearity.

Is the assumption valid? **Yes**

CLM 4 - Zero conditional mean

We'll now plot our model in order to assess if the model has zero conditional mean.

```
plot(model1, which=1)
```



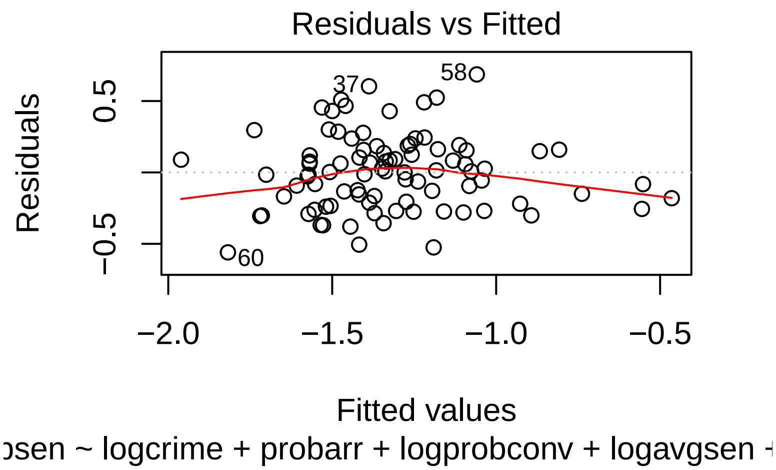
The red line is staying relatively close to the X-axis for the most part although it is influenced by the outliers on the ends.

Is the assumption valid? **Yes**

CLM 5 - Homoscedasticity

We will use the same plot to assess the model's homoscedasticity.

```
plot(model1, which=1)
```



The plot is relatively scattered about the fitted values with some extreme outliers. It is a little bit difficult to determine if we have achieved homoscedasticity from this plot alone. We will run a couple of additional tests to determine the homoscedasticity of the model.

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 4.0657, df = 6, p-value = 0.6678
```

```
ncvTest(model1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.6267145    Df = 1     p = 0.428563
```

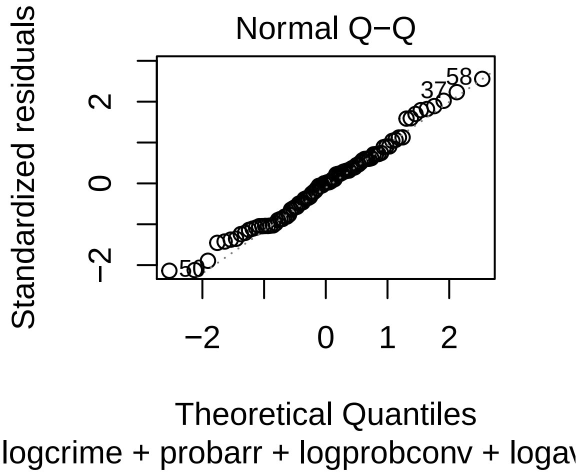
Neither test is showing a small enough P-value suggesting we fail to reject the null hypothesis of homoscedasticity. Therefore we most likely have homoscedasticity however looking at the plot, it is a little bit questionable.

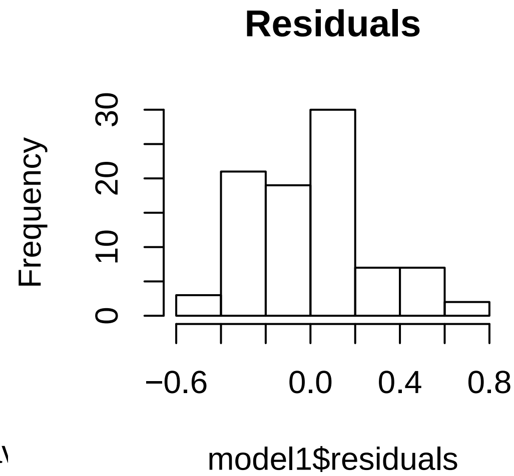
Is the assumption valid? **Most likely**

CLM 6 - Normality of residuals

We will now lok at the QQ-plot to assess the normality of residuals.

```
plot(model1, which=2)
hist(model1$residuals, main="Residuals")
```



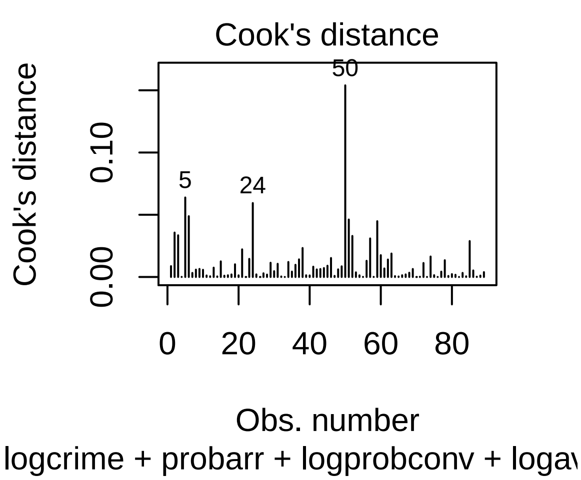


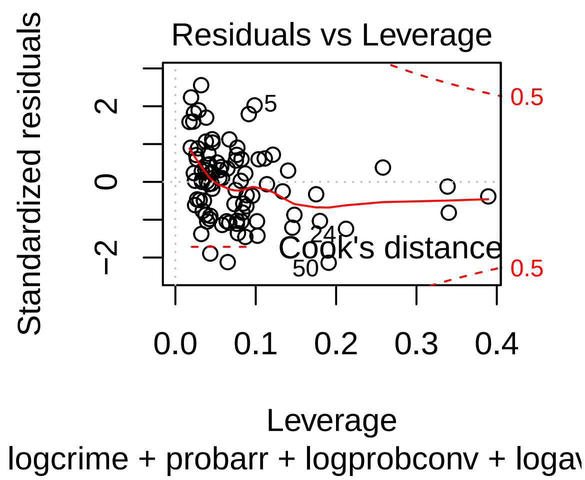
The values are staying close to the slope for the most part however are deviating on both ends. However the distribution of the residuals is relatively normal and our sample size n is 90 and hence CLM 6 is achieved.

Is the assumption valid? **Yes**

Cook's distance

```r
plot(model1, which = 4)
plot(model1, which = 5)
```





There is a influencial value at 50 however it is still well within the bounds of Cook's distance.

AIC

```r
(model1$AIC <- AIC(model1))
```

```
## [1] 30.08571
```

The AIC for this model is 30.0857125.

Propposed Model #2 - Optimal specification

In addition to the set of explanatory variables introduced in Proposed Model #1, we have decided to include the following variables in this model:

Demographics variables: The team is interested to see if demographics information such as race, gender and age would influence the probability of prison sentence. We are including *pctmin*, *ymale* in this model to assess this.

Density: The team suspects population density would have a negative influence on *probsen* by introducing more complexity in crimes.

Tax: The team anticipate *tax* would have a negative coefficient as higher tax revenue usually suggests people have more money. People with more money are typically able to afford better lawyers and hence would have lower chances of ending up with prison sentences.

$$log(probsen) = \beta_0 + \beta_1 log(crime) + \beta_2 probarr + \beta_3 log(probconv) + \beta_4 log(avgsen) + \beta_5 log(police)$$

$$+\beta 6 log(density) + \beta 7 log(tax) + \beta 8 log(pctmin) + \beta 9 log(mix) + \beta 10 log(ymale) + u$$

```r
model2 <- lm(logprobsen ~ logcrime + probarr + logprobconv + logavgsen + logpolice
             + logdensity + logtax + logpctmin + logmix + logymale, data = D)
```

CLM

No change in CLM1-2.

CLM 3 - Multicollinearity

We'll compute VIF

```r
vif(model2)
```

```
##    logcrime     probarr logprobconv   logavgsen   logpolice  logdensity
##    3.724762    1.070785    1.796106    1.159418    2.126302    2.456174
##      logtax   logpctmin      logmix    logymale
##    1.540660    1.809699    1.712168    1.318453
```

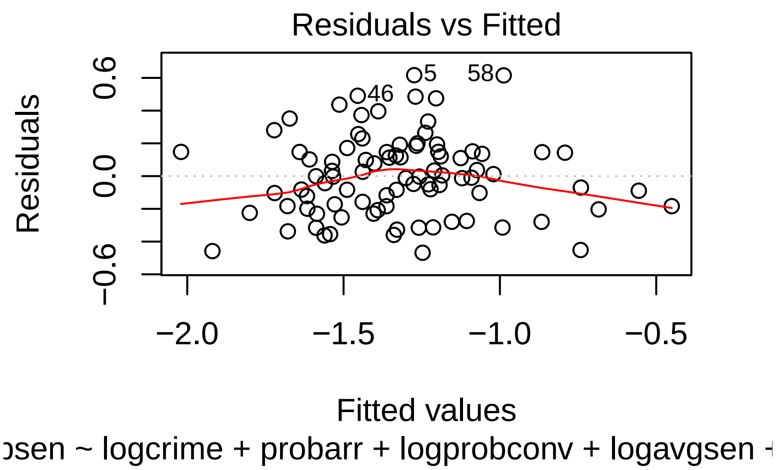
All computed VIF values are < 4.

Is the assumption valid? **Yes**

CLM 4 - Zero conditional mean

We'll now plot our model in order to assess if the model has zero conditional mean.

```r
plot(model2, which=1)
```



The fitted line is staying relatively close to the X-axis for the most part however is influenced by the outliers on the both sides.

Is the assumption valid? **Yes**

CLM 5 - Homoscedasticity

The plot is relatively distributed evenly about the fitted values.

```r
bptest(model2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model2
## BP = 12.476, df = 10, p-value = 0.2544
```

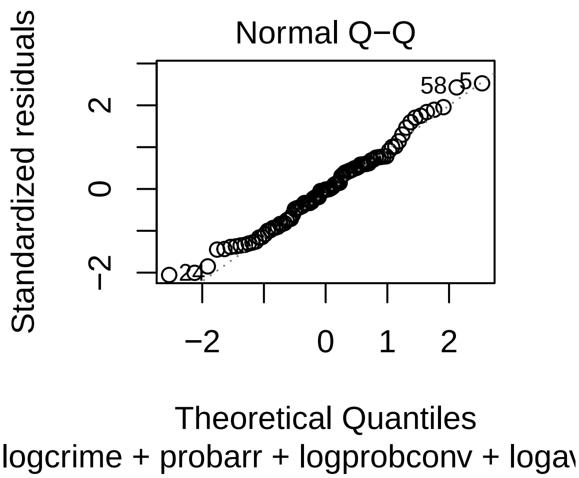
Checking the BP test result, the P-value is not small enough to reject the null hypothesis of homoscedasticity.

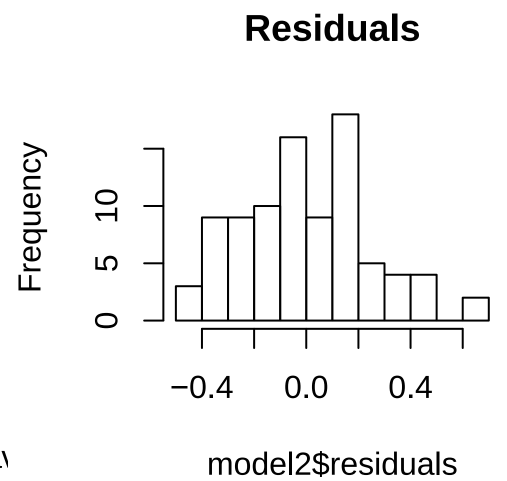
Is the assumption valid? **Most likely**

CLM 6 - Normality of residuals

We will now lok at the QQ-plot to assess the normality of residuals.

```r
plot(model2, which=2)
hist(model2$residuals, main = "Residuals")
```



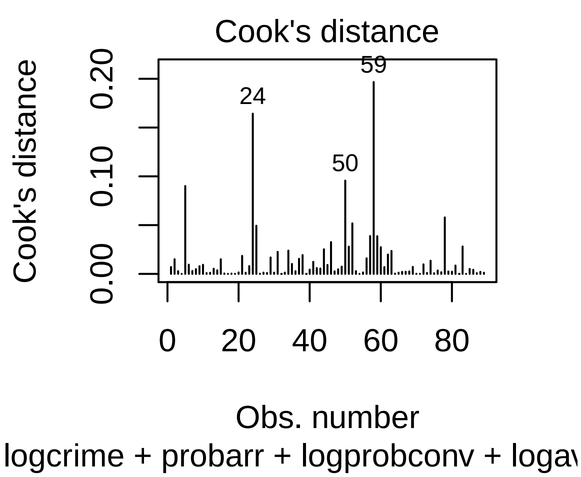


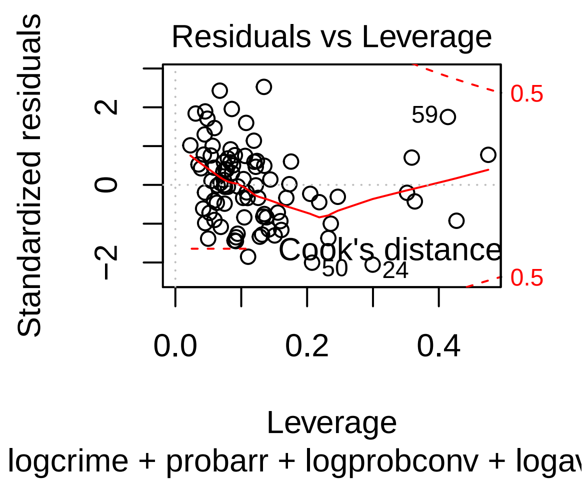
The both plots are showing we have normality of residuals.

Is the assumption valid? **Yes**

Cook's distance

```r
plot(model2, which = 4)
plot(model2, which = 5)
```





There is a influencial value at 59 however it is still well within the bounds of Cook's distance.

AIC

```r
(model2$AIC <- AIC(model2))
```

```
## [1] 26.70092
```

The AIC for this model is 26.7009226 which is lower compared to model 1 indicating this is an improved model.

Proposed Model 3 - Comprehensive specification

This model includes all variables present in the dataset to show the robustness of my modeling process and the underlying assumptions to model specification.

$$log(probsen) = \beta_0 + \beta_1 log(crime) + \beta_2 probarr + \beta_3 log(probconv) + \beta_4 log(avgsen) + \beta_5 log(police)$$

$$+\beta 6 log(density) + \beta 7 log(tax) + \beta 8 log(pctmin) + \beta 9 log(mix) + \beta 10 log(ymale)$$

$$+\beta 11 west + \beta 12 central + \beta 13 urabn + \beta 14 log(wagecon) + \beta 15 wagetuc + \beta 16 wagetrd$$

$$+\beta 17 wagefir + \beta 18 log(wageser) + \beta 19 wagemfg + \beta 20 wagefed + \beta 21 wagesta + \beta 22 wageloc + u$$

```
model3 <- lm(logprobsen ~ logcrime + probarr + logprobconv + logavgsen + logpolice
             + logdensity + logtax + logpctmin + logmix + logymale + west + central + urban
             + logwagecon + wagetuc + wagetrd + wagefir + logwageser + wagemfg + wagefed
             + wagesta + wageloc, data = D)
```

CLM

No change in CLM1-2.

CLM 3 - Multicollinearity

We'll compute VIF

```
vif(model3)
```

```
##    logcrime     probarr logprobconv   logavgsen   logpolice  logdensity
##    4.863753    1.174475    2.054750    1.557634    2.889421    6.081971
##      logtax   logpctmin      logmix    logymale        west     central
##    2.455541    4.448054    2.071487    1.698017    2.267788    4.878311
##       urban  logwagecon     wagetuc     wagetrd     wagefir  logwageser
##    2.928001    2.186521    1.745930    3.228493    2.931556    1.663908
##     wagemfg     wagefed     wagesta     wageloc
##    2.008753    3.522783    1.699442    2.354271
```

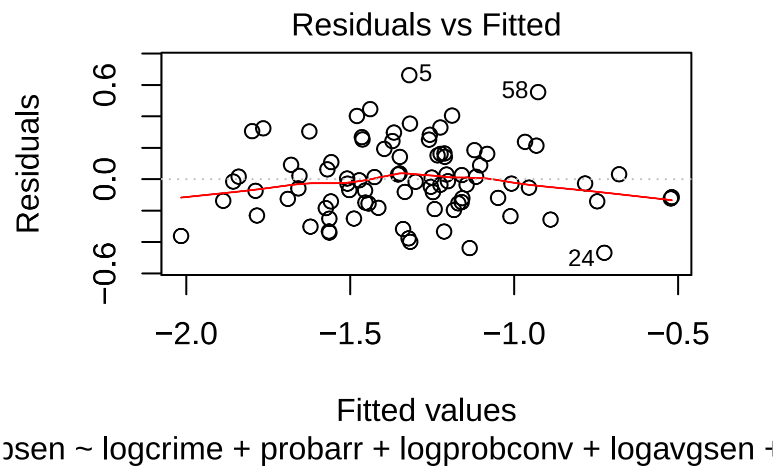
All the values are < 10.

Is the assumption valid? **Yes**

CLM 4 - Zero conditional mean

We'll now plot our model in order to assess if the model has zero conditional mean.

```
plot(model3, which=1)
```



The fitted line is staying relatively close to the X-axis for the most part.

Is the assumption valid? **Yes**

CLM 5 - Homoscedasticity

The plot is relatively distributed evenly about the fitted values.

```r
bptest(model3)
```

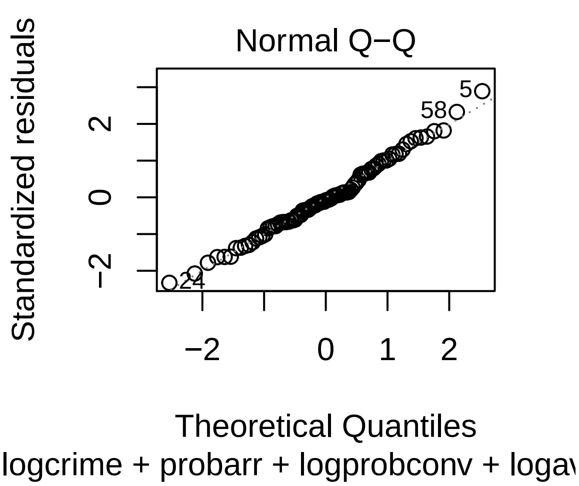
```
##
##  studentized Breusch-Pagan test
##
## data:  model3
## BP = 25.311, df = 22, p-value = 0.2825
```

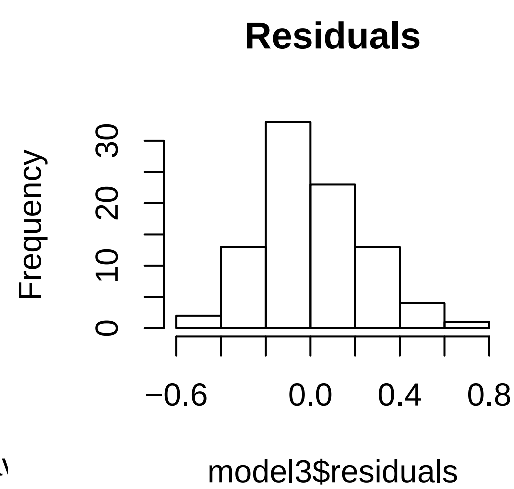
Checking the BP test result, the P-value is not small enough to reject the null hypothesis of homoscedasticity.

Is the assumption valid? **Most likely**

CLM 6 - Normality of residuals

```r
plot(model3, which=2)
hist(model3$residuals, main = "Residuals")
```



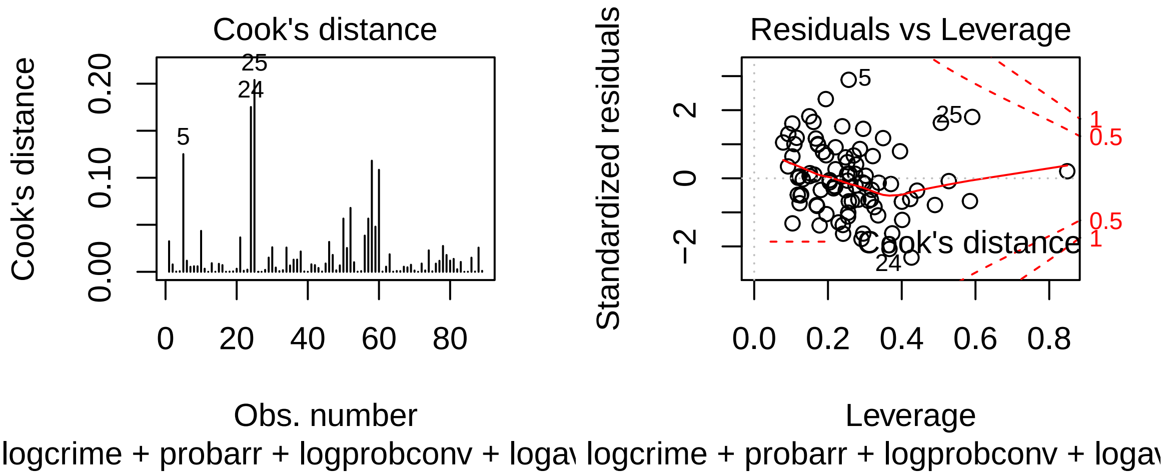


The both plots are showing we have normality of residuals.

Is the assumption valid? **Yes**

Cook's distance

```r
plot(model3, which = 4)
plot(model3, which = 5)
```



There are some spikes however they are still well within the bounds of Cook's distance.

AIC

```r
(model3$AIC <- AIC(model3))
```

```
## [1] 37.99468
```

The AIC for this model is 37.9946771 which is the highest of the 3 models, suggesting this is not a very good model according to AIC.

Model Adjustments

We will now be adjusting the models in order as there were some CLM assumptions that were violated or not entirely met.

CLM 4 - Zero conditional mean

The fitted value plots for some of our models showed curvature towards the ends most likely influenced by outliers. Since our sample size n is relatively large, we may be able to use MLR 4' Zero mean and zero correlation (exogenity) instead of the standard CLM 4 assumption in order to address this violation.

CLM 5 - Homoscedasticity

In order to address the possible violations of CLM 5 Homoscedasticity assumption, we will calculate heteroscedasticity robust standard errors.

Model 1

```r
coef(summary(model1))[, 2]
```

```
## (Intercept)    logcrime     probarr logprobconv    logavgsen    logpolice
##  0.79178131  0.06894510  0.36987910  0.06209656  0.11618225  0.11195755
##      logmix
##  0.05881954
```

```r
(model1$se.adjusted <- sqrt(diag(vcovHC(model1))))
```

```
## (Intercept)    logcrime     probarr logprobconv    logavgsen    logpolice
##  0.67896991  0.08267105  0.34124047  0.06952464  0.10253441  0.09683524
##      logmix
##  0.06779809
```

Model 2

```r
coef(summary(model2))[, 2]
```

```
## (Intercept)    logcrime     probarr logprobconv    logavgsen    logpolice
##  1.12072823  0.10331934  0.35931329  0.06841172  0.11212742  0.12511552
##  logdensity      logtax   logpctmin      logmix     logymale
##  0.05627013  0.13101419  0.04084239  0.06643983  0.16132430
```

```r
(model2$se.adjusted <- sqrt(diag(vcovHC(model2))))
```

```
## (Intercept)    logcrime     probarr logprobconv    logavgsen    logpolice
##  1.43429852  0.12265221  0.38445249  0.08437148  0.10627715  0.13583936
##  logdensity      logtax   logpctmin      logmix     logymale
##  0.05989955  0.14400446  0.04552969  0.07354445  0.24169374
```

Model 3

```r
coef(summary(model3))[, 2]
```

```
##  (Intercept)     logcrime      probarr  logprobconv    logavgsen
## 2.0375701946 0.1195066365 0.3809061069 0.0740658421 0.1315522760
##    logpolice   logdensity       logtax    logpctmin       logmix
## 0.1476311904 0.0896281438 0.1674219738 0.0648137013 0.0739725204
##     logymale         west      central        urban   logwagecon
## 0.1853157460 0.0872669332 0.1489925324 0.1684485986 0.2593767549
##      wagetuc      wagetrd      wagefir   logwageser      wagemfg
## 0.0004848938 0.0014940057 0.0008935393 0.1260606829 0.0004565242
##      wagefed      wagesta      wageloc
## 0.0008815037 0.0008485928 0.0015995764
```

```r
(model3$se.adjusted <- sqrt(diag(vcovHC(model3))))
```

```
##  (Intercept)     logcrime      probarr  logprobconv    logavgsen
## 2.3887671032 0.1904656805 0.4114521222 0.1131892621 0.1429948150
##    logpolice   logdensity       logtax    logpctmin       logmix
## 0.1650621562 0.1043631263 0.2305953809 0.0901974581 0.0924087356
##     logymale         west      central        urban   logwagecon
```

```
## 0.2796174017 0.1064263357 0.1649583437 0.1931151236 0.2744396082
##      wagetuc      wagetrd      wagefir   logwageser      wagemfg
## 0.0006598016 0.0017896417 0.0011538875 0.1111826014 0.0004584581
##      wagefed      wagesta      wageloc
## 0.0010681425 0.0010781599 0.0025036122
```

Heteroscedasticity robust standard errors tend to be more conservative. You can confirm by looking at the values of the robust standard errors which tend to be larger than those of the original standard errors.

Model Analysis

```r
stargazer(model1, model2, model3, omit.stat = "f", header=FALSE,
          title = "Models for predicting probability of prison sentences",
          se =
            list(model1$se.adjusted, model2$se.adjusted,
                 model3$se.adjusted),
          star.cutoffs = c(0.05, 0.01, 0.001), no.space = TRUE)
```

Table 1: Models for predicting probability of prison sentences

	Dependent variable:		
	logprobsen		
	(1)	(2)	(3)
logcrime	-0.304^{***}	-0.412^{***}	-0.498^{**}
	(0.083)	(0.123)	(0.190)
probarr	-0.326	-0.359	-0.314
	(0.341)	(0.384)	(0.411)
logprobconv	-0.201^{**}	-0.305^{***}	-0.327^{**}
	(0.070)	(0.084)	(0.113)
logavgsen	-0.031	-0.025	-0.021
	(0.103)	(0.106)	(0.143)
logpolice	-0.051	0.062	0.013
	(0.097)	(0.136)	(0.165)
logdensity		0.023	0.088
		(0.060)	(0.104)
logtax		-0.073	0.071
		(0.144)	(0.231)
logpctmin		0.105^{*}	0.177^{*}
		(0.046)	(0.090)
logmix	0.310^{***}	0.214^{**}	0.217^{*}
	(0.068)	(0.074)	(0.092)
logymale		-0.336	-0.193
		(0.242)	(0.280)
west			-0.044
			(0.106)
central			0.157
			(0.165)
urban			-0.205
			(0.193)
logwagecon			0.211
			(0.274)
wagetuc			-0.0001
			(0.001)
wagetrd			-0.00000
			(0.002)
wagefir			-0.001
			(0.001)
logwageser			-0.091
			(0.111)
wagemfg			-0.00003
			(0.0005)
wagefed			0.001
			(0.001)
wagesta			-0.001
			(0.001)
wageloc			-0.0001
			(0.003)
Constant	-1.991^{**}	-2.813^{*}	-4.492
	(0.679)	(1.434)	(2.389)
Observations	89	89	89
R^2	0.511	0.570	0.627
Adjusted R^2	0.475	0.515	0.503
Akaike Inf. Crit.	30.086	26.701	37.995
Residual Std. Error	0.273 (df = 82)	0.262 (df = 78)	0.266 (df = 66)
Note:			*p<0.05; **p<0.01; ***p<0.001

Model 1

logcrime, *logmix* and *logprobconv* have very small P-values suggesting strong stastical significance. Contrary to my initial hypothesis, all the original co-efficients except for logmix are negative. 1% increase in *logcrime* and *logprobconv* results in -30.4% and -20.1% impact on the dependent variable *probsen* which are both practically significant.

Adjusted R^2 is 0.475 which is the lowest of the 3 models, explanining 47.5% of the variation in *log(probsen)*.

Model 2

In addition to *logcrime*, *logmix*, *logprobconv* and *logpctmin* has a P-value < 0.05 in this model. It has a positive coefficient indicating in 1% increase in *logpctmin* will translate into 10.5% increase in *probsen* which is a practically significant result.

Adjusted R^2 is 0.515 which is the highest of the 3 models, explanining 51.5% of the variation in *log(probsen)*. The model also has the lowest AIC of the 3 models at 26.701 indicating this is the best model of the 3 according to Akaike's Information Criterion.

Model 3

The same set of variables as Model 2, *logcrime*, *logmix*, *logprobconv* and *logpctmin* are showing statistical significance although not as strongly. One thing to note is that the co-efficient values for the statistically significant covariates in this model appear to be larger in the magnitude and hence practical significance than those of Model 2. For example, *logcrime* is showing -0.498 which is greater vs -0.412 for Model 2.

Adjusted R^2 is 0.503 which is the second highest of the 3 models, explanining 50.3% of the variation in *log(probsen)*. The model also has the highest AIC of the 3 models at 37.995 indicating this is the worst model of the 3 according to Akaike's Information Criterion.

Causality

Our current models account for roughly 50% of variance in the dependent variable *probsen* with the following variables having the most prominent influence: *logcrime*, *logmix*, *logprobconv* and *logpctmin*.

Interestingly, the 2 most statistically significant explanatory variables *logcrime* and *logprobconv* have negative coefficients suggesting that increase in these variables would result in a decrease in *logprobsen*. Some of the possible causes of this maybe overcrowding of the local prisons or the local judical system and police force being overworked and not working effectively. We are unable to determine what is causing this seemingly counterintuitive phenemenon from the data given.

On the other hand, *logmix* and *logpctmin* have positive coefficients suggesting increase in them would yield higher *probsen*. It is easy to see why *mix*, the ratio of face to face/all other crimes, would produce this result as face-to-face crimes are clearly easier to prosecute. Although it is not as statistically nor practically significant, *logpolice* also has a positive coefficient suggesting more eyes in the field may produce safer communities. The fact *logpctmin* has a positive coefficient suggests that there may be a prejudice in our judical system that are biased against people of certain races or people of certain races are more likely to be involved in serious crimes that end up in prison sentences.

Omitted variable bias

For this research, since we were given a set of variables to work with in a CSV file rather than identifying and collecting relevant data ourselves, it is quite possible we have a case of omitted variable bias. For example, our original dataset did not contain any data on poverty rate which may have been useful. The wage* variables are informative in learning the economic standings of those that work in a given sector, however, they do not tell anything about those without jobs who may be involved in crimes. Also, the dataset did not contain much demographic data other than *pctmin* and *ymale* which report the proportion that is minority or nonwhite and} and the proportion of county males between the ages of 15-24 respectively. More comprehensive demographic data such as more comprehensive age and gender data, ratio of immigrants and people's educational backgrounds may have been resourceful in designing a more exhaustive model.

Selection bias

In the dataset included in crime.csv, there were 90 rows each representing a county. The data was collected, hand-picked and given to us for the purpose of the research and we assumed that the dataset represents a fair representation of the relevant counties for the political campaign. However, we cannot deny the possibility that there may have been a selection bias in choosing which counties to include.

Conclusion

Our models suggest that face-to-face crimes *logmix* and minorities *logpctmin* are amongst the key positively contributing factors in the variance of our dependent variable *logprobsen*. More patrols and surveillance cameras in the areas where moniroity population is predominant may help mitigate the number of offenses that will end up in prison sentences.

Another interesting finding was that the counties with high crime and conviction rates actually have a lower probability of prison sentence. This may be due to the fact that the judical and system and the police force in the counties with high *crime* and {emph{probconv} are working at full capacity and are not able to conduct through investigations failing to land cases in prison sentences. It could also be the case that the prisons in the area may be simply full due to the high crime rate and are not accepting as many incoming inmates. A more funding to research the current state of the local judical system including the prisons and the police force may be a good starting point in shedding light on this phenomenon.

Based on these, our proposals for the campaign are as follow:

1. Increase the number of patrols by the police and surveillance cameras in the communities with high face-to-face crime rates with an emphasis on nonwhite neighborhoods
2. Conduct a study on the counties with high crime rates and conviction rates and why the crimes committed in these counties are not resulting in as many prison sentences