Lab 4: Reducing Crime

Kim Vignola, Kiersten Hendersen, Aaron Yuen

August 17, 2017

Section 1: Introduction

The authors, Kim, Kiersten and Aaron, were hired to provide research for a political campaign in North Carolina to understand the determinants of crime (both correlational and causal) using exploratory data analysis and OLS regression. The end goal is to leverage the data to provide policy suggestions that are applicable to local government to reduce crime.

The provided dataset consists of statistics for a selection of counties for a given time period. Data for 90 counties and 25 variables for each county were provided.

For this analysis, the following assumptions were made:

- The 90 counties provided were randomly sampled among the 100 counties in North Carolina.

Section 2: Exploratory Analysis

Transformed Dataset

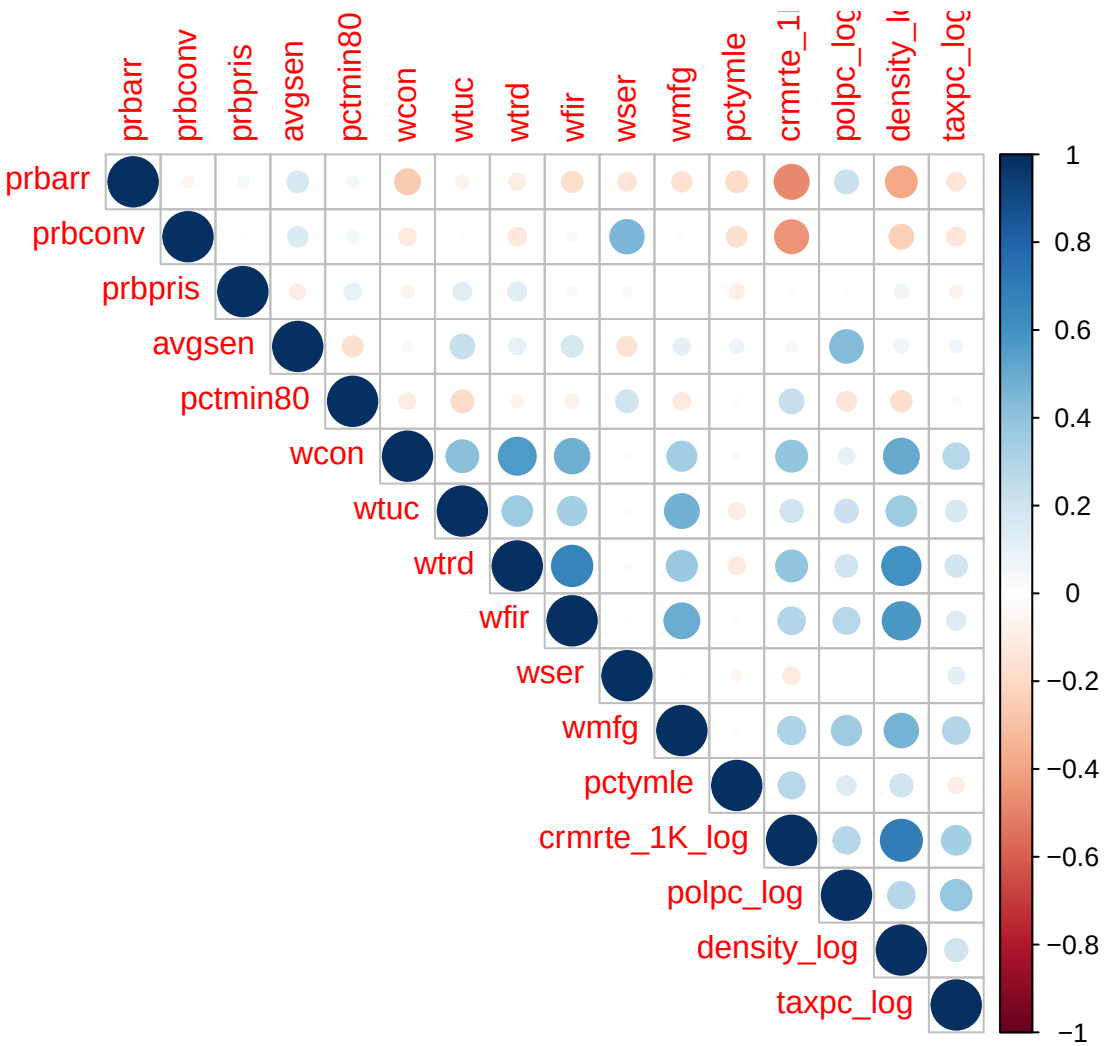
```
library(corrplot)
library(car)
library(lmtest)
library(sandwich)

setwd("C:\\Users\\aayuen\\Documents\\GitHub\\w203_lab4_kka")
data = read.csv("crime.csv")
data$crmrte_1K_log = log(data$crmrte *10^3)
data$polpc_log = log(data$polpc)
data$density_log = log(data$density)
data$taxpc_log = log(data$taxpc)
data$pctmin80 = data$pctmin80 / 100
data$wtotal = (data$wcon + data$wtuc + data$wtrd + data$wfir +
               data$wser + data$wmfg + data$wfed + data$wsta + data$wloc)
data = data[data$X != 81,]

sort(colnames(data))
```

```
##  [1] "avgsen"        "central"       "county"        "crmrte"
##  [5] "crmrte_1K_log" "density"       "density_log"   "mix"
##  [9] "pctmin80"      "pctymle"       "polpc"         "polpc_log"
## [13] "prbarr"        "prbconv"       "prbpris"       "taxpc"
## [17] "taxpc_log"     "urban"         "wcon"          "west"
## [21] "wfed"          "wfir"          "wloc"          "wmfg"
## [25] "wser"          "wsta"          "wtotal"        "wtrd"
## [29] "wtuc"          "X"             "year"
```

Analysis on Dataset

MORE TO EXPLAIN HERE

```
corrplot(cor(data[ , names(data) %in%
                     c("crmrte_1K_log", "density_log", "avgsen", "prbarr", "prbconv", "prbpris", "polpc_log",
            "taxpc_log", "pctmin80", "pctymle", "density_1K_log",
            "wcon", "wtuc", "wtrd", "wfir", "wser", "wmfg")]),
         method="circle", type="upper")
```



Section 3: Model Building and Assumptions

Model #1 - With Explanatory Variables of Key Interest

Based on the exploratory analysis, finding key independent variables with considerable correlation with log of crime rate, we will first look at log of density, probability of arrest, probability of conviction, log of tax revenue per capita and percent young male.

MORE TO EXPLAIN HERE

```
m1 = lm(crmrte_1K_log ~ density_log + prbarr
        + prbconv + taxpc_log + pctymle, data=data)
```

MLR.1 - Linear in Parameters

The model is speci???ed such that the dependent variable is a linear function of the explanatory variables.

As a result, MLR.1 holds true.

MLR.2 - Random Sampling

There are only 90 counties in the dataset, and 1 was removed due to it being a suspected outlier, while there are a total of 100 counties in North Carolina since 1911 (and data indicates it is from 1987). As a result, the dataset does *not* contain all of 100 counties in North Carolina. That said, there is no obvious pattern to suggest that there was any concious decision to keep or remove certain counties.

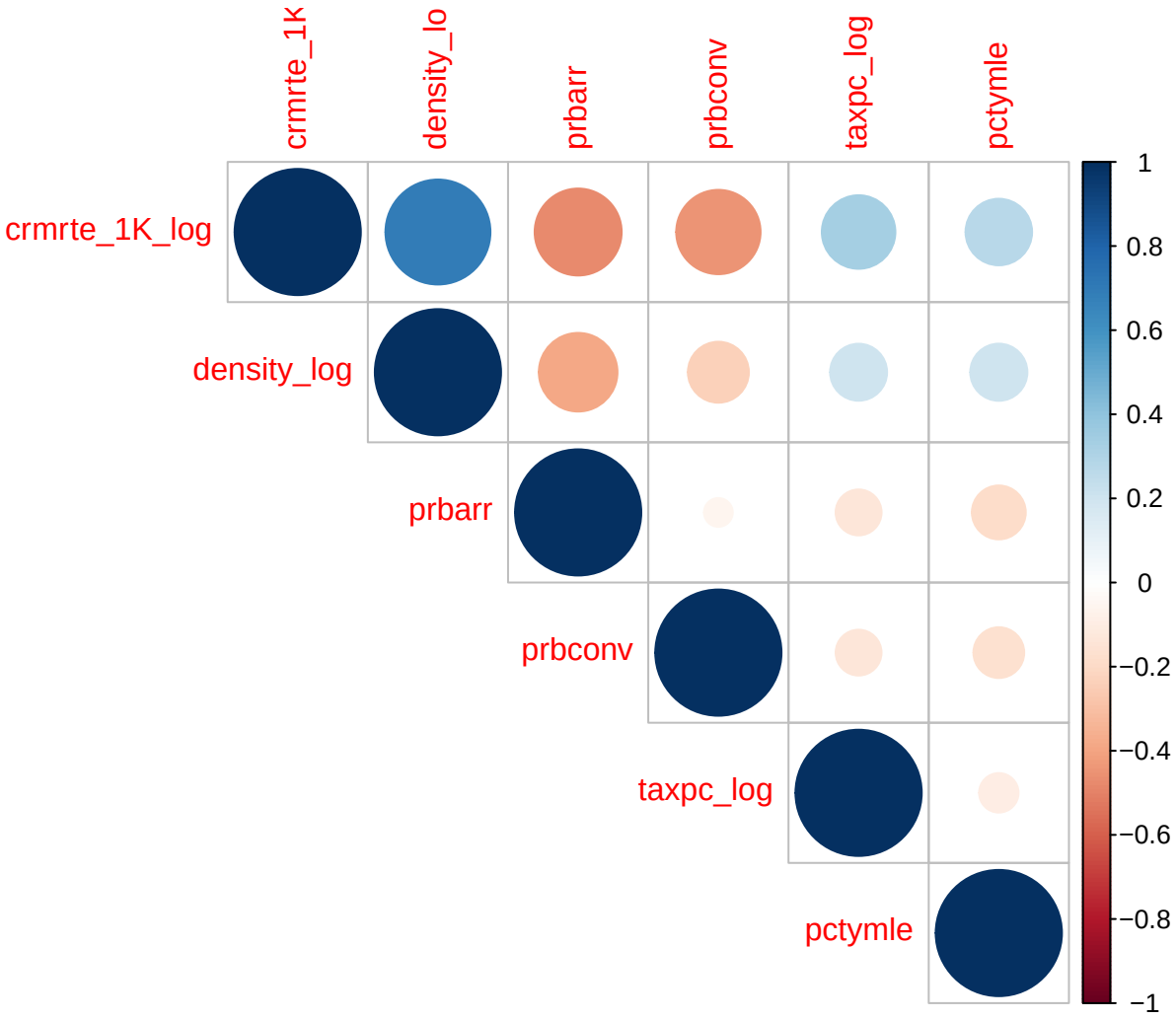
As a result, for this analysis we assume that the counties were randomly sampled from the 100 counties in North Carolina.

As a result, MLR.2 holds true.

MLR.3 - Multicollinearity

From the correlation plot, there is no evidence of perfectly correlated variable pairs.

```
X = data.matrix(subset(data, select=c("crmrte_1K_log", "density_log", "prbarr", "prbconv", "taxpc_log", "pcty
corrplot(cor(X), method="circle", type="upper")
```



Looking at the VIFs, they all are less than 10, which suggests that ther eis no perfect multicollinearity of the independent variables.

```
vif(m1)
```

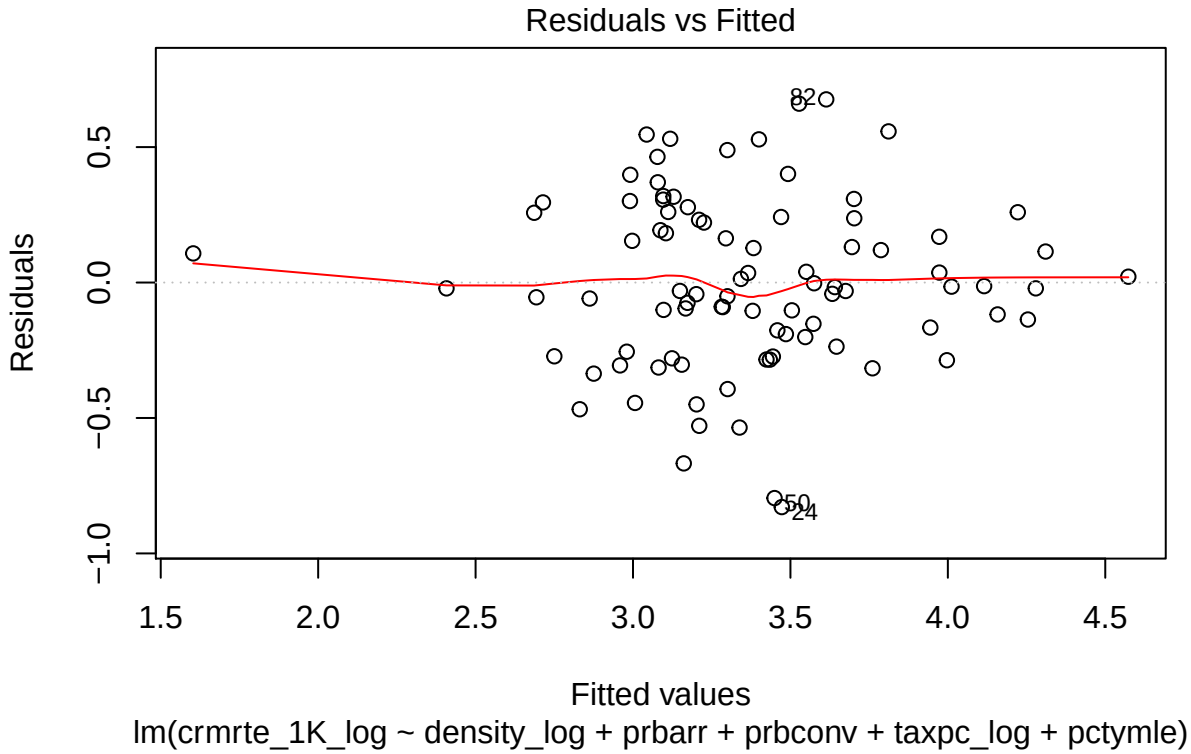
```
## density_log      prbarr     prbconv   taxpc_log     pctymle
##    1.319348    1.242287    1.123784    1.088815    1.110903
```

As a result, MLR.3 holds true.

MLR.4 - Zero-Conditional Mean

The residuals vs fitted plot indicates little evidence that that the zero-conditional mean assumption doesn't hold. For example, the red spline is mostly along 0.

```
plot(m1, which=1)
```



Next, looking at the covariances of the independent variables with the residuals, they all are very close to zero, indicating they are likley exogenous.

```
cov(data$density_log, m1$residuals)
```

```
## [1] -8.624643e-18
```

```
cov(data$prbarr, m1$residuals)
```

```
## [1] 3.441283e-18
```

```
cov(data$prbconv, m1$residuals)
```

```
## [1] -3.54568e-18
```

```
cov(data$taxpc_log, m1$residuals)
```

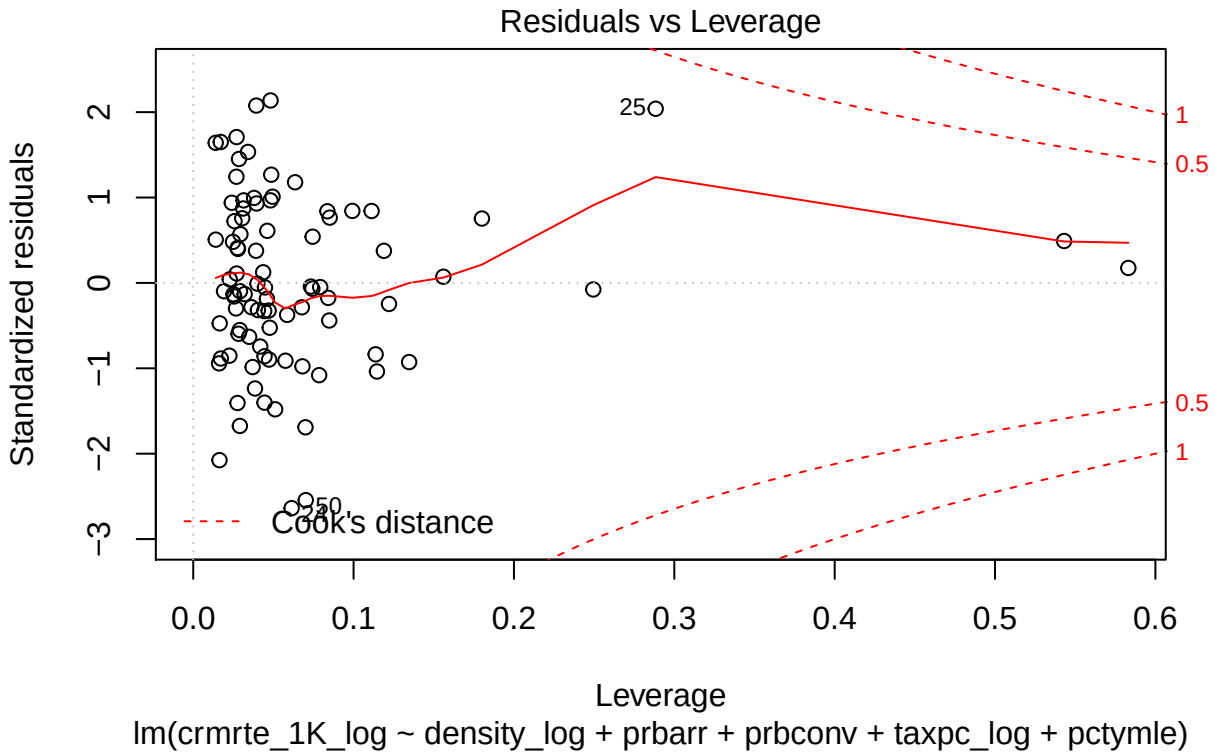
```
## [1] -7.399218e-19
```

```
cov(data$pctymle, m1$residuals)
```

```
## [1] -4.207396e-19
```

Lastly, there are no data points with a large Cook's distance, so there is likely no observations with undue influence on the model fit.

```
plot(m1, which=5)
```

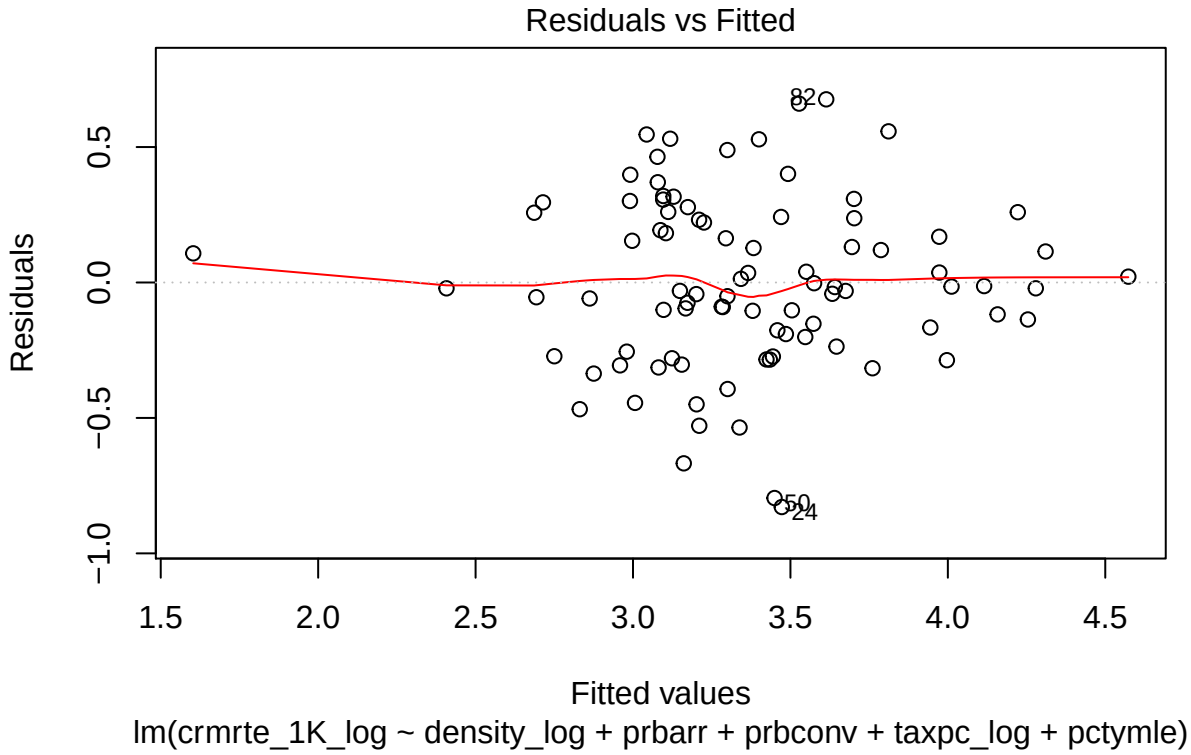


As a result, MLR.4 holds true.

MLR.5 - Homoscedasticity

Going back to the residuals vs fitted plot, it looks like the variance of errors to the right of the plot is not constant with those in the middle of the plot.

```
plot(m1, which=1)
```



To further test this, the Breusch-Pagan test, which the null hypothesis is there is homoskedasticity, has a low p-value. This suggests there is a heteroscedasticity problem.

```
bptest(m1)
```

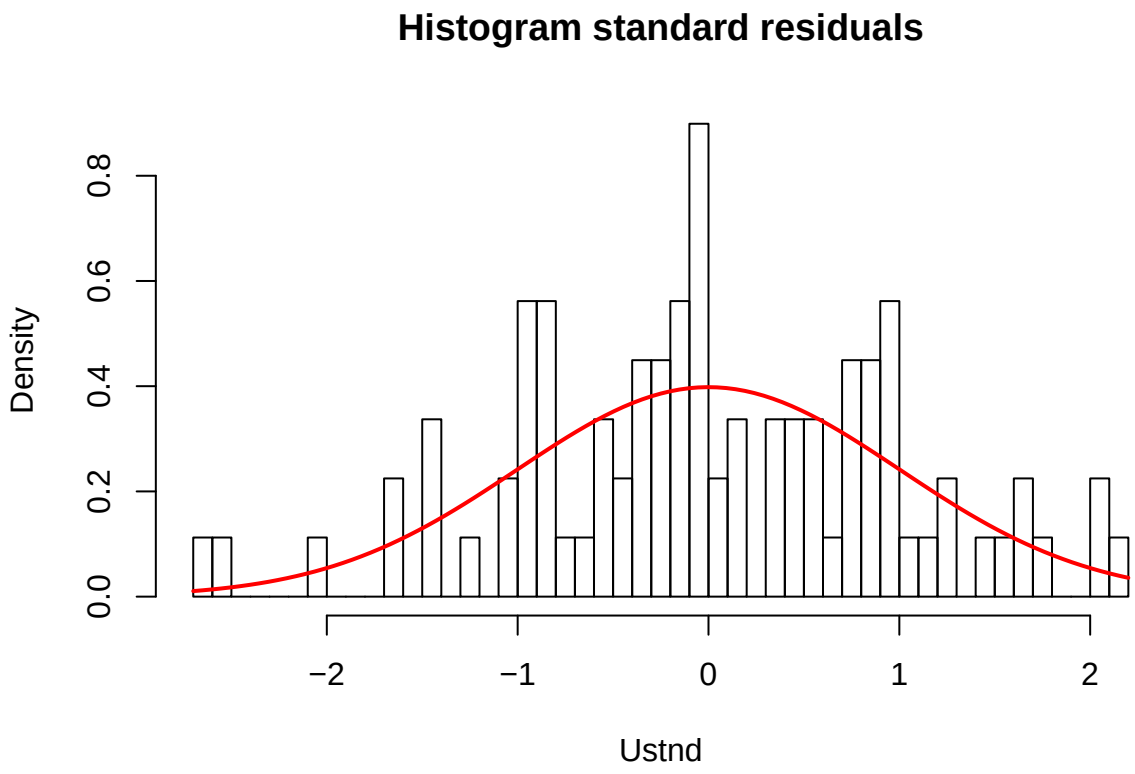
```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 17.738, df = 5, p-value = 0.003293
```

As a result, MLR.5 does *not* hold true. To address this, we will be using robust standard erros when analyzing the model statistics.

MLR.6 - Normality of Residuals

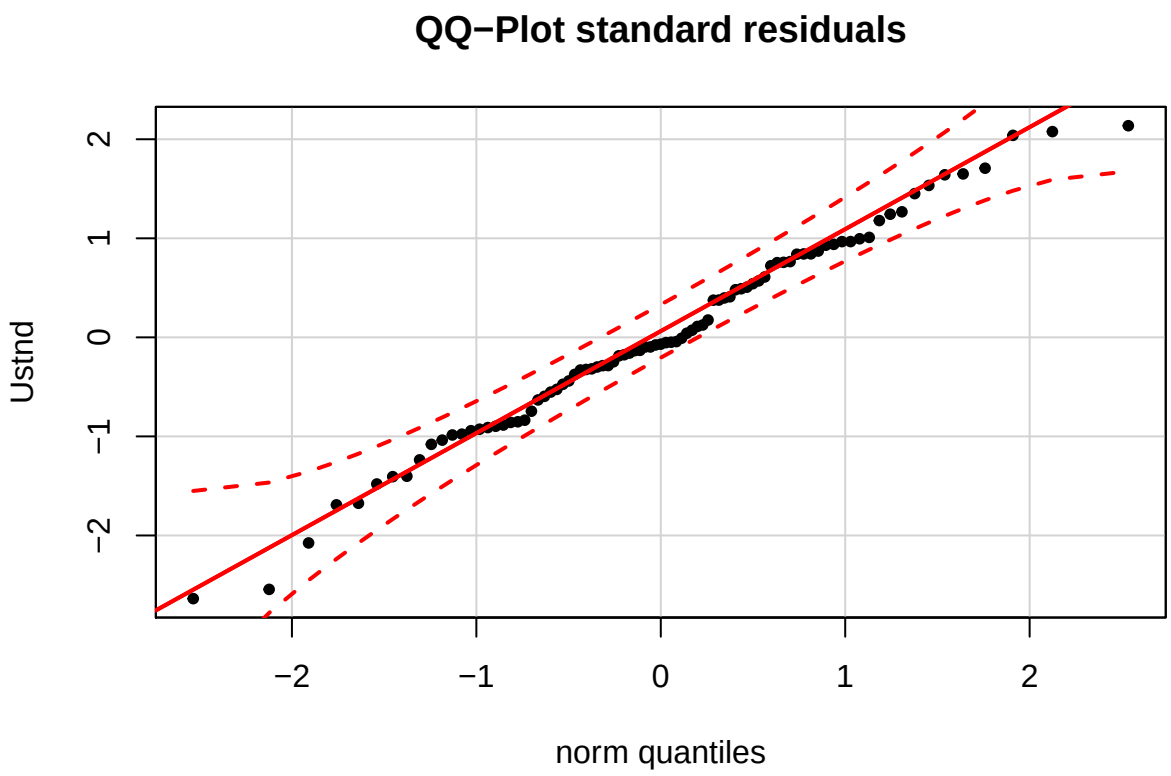
Looking at the histogram of the standard residuals, it doesn't look like the residuals are clearly normal.

```
Ustnd = rstandard(m1)
hist(Ustnd, main="Histogram standard residuals", breaks = 50, freq=FALSE)
curve(dnorm(x, mean=0, sd=sd(Ustnd)), col="red", lwd=2, add=TRUE)
```



Further confirming with the QQ-plot, there is evidence to suggest that the residuals does not closely follow a normal distribution.

```
qqPlot(Ustnd, distribution="norm", pch=20, main="QQ-Plot standard residuals")
qqline(Ustnd, col="red", lwd=2)
```



As a result, MLR.6 does *not* hold true. That said, given that n is considerably > 30, we can rely on asyptotic properties of OLS.

Model statistics

To adust the violated assumptions on MLR.5, we will use robust standard errors for looking at the model statistics.

Based on the coeftest, there are a few coefficients that are statistically significant:

1. Intercept - with p-value 0.0013
2. density_log - with p-value 1.34e-07
3. prbarr - with p-value 0.00015
4. prbconv - with p-value 9.27e-06
5. pctymle - with p-value 0.021

```
coeftest(m1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.413434   0.724801  3.3298 0.0012982 **
## density_log  0.319888   0.055464  5.7674 1.339e-07 ***
## prbarr      -1.087047   0.274153 -3.9651 0.0001548 ***
## prbconv     -0.488206   0.103323 -4.7250 9.272e-06 ***
## taxpc_log    0.371339   0.194745  1.9068 0.0600066 .
## pctymle      2.380534   1.014963  2.3454 0.0213911 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model #2 - With Explanatory Variables of Key Interest and Potential Covariates to Increase Accuracy

MORE TO ADD LATER

Model #3 - With Previous Covariates and Most Other Covariates

MORE TO ADD LATER

Section 4: Model Summary

MORE TO ADD LATER

Section 5: Causality

MORE TO ADD LATER

Section 6: Conclusion

MORE TO ADD LATER