ECON 104 Project 1

Marc Luzuriaga, Takuya Sugahara, Daniel Day, Shabib Alam

2023-10-14

```r
#Loading in the MurderRates Data Set
data("MurderRates")
```

1 Introduction: Murder Rate Determinants

1.1 Data Set Summary

In this paper, we will be analyzing the Murder Rate Data Set from the AER Package. The data set holds cross-sectional data on states in the year 1950, and the data set contains 44 observations on the following 8 variables:

(1) rate: Murder rate per 100,000 (FBI estimate, 1950)

(2) convictions: Number of convictions divided by number of murders in 1950.

(3) executions: Average number of executions during 1946-1950 divided by convictions in 1950.

(4) time: Median time served (in months) of convicted murderers released in 1951.

(5) income: Median family income in 1949 (in 1,000 USD).

(6) lfp: Labor force participation rate in 1950 (in percent).

(7) noncauc: Proportion of population that is non-Caucasian in 1950.

(8) Southern: Factor indicating region

1.2 Question

The question we seek to answer with the Murder Rates Data Set is as follows: "Does the Median family income in 1949 (in 1,000 USD) or the median time served (in months) of convicted murderers released in 1951 have an equal effect in reducing murder rates in states?" For the purposes of this project, our group will claim that the median family income in 1949 (in 1,000 USD) and an increased median time served in prison will have an equal effect on decreasing murder rates.

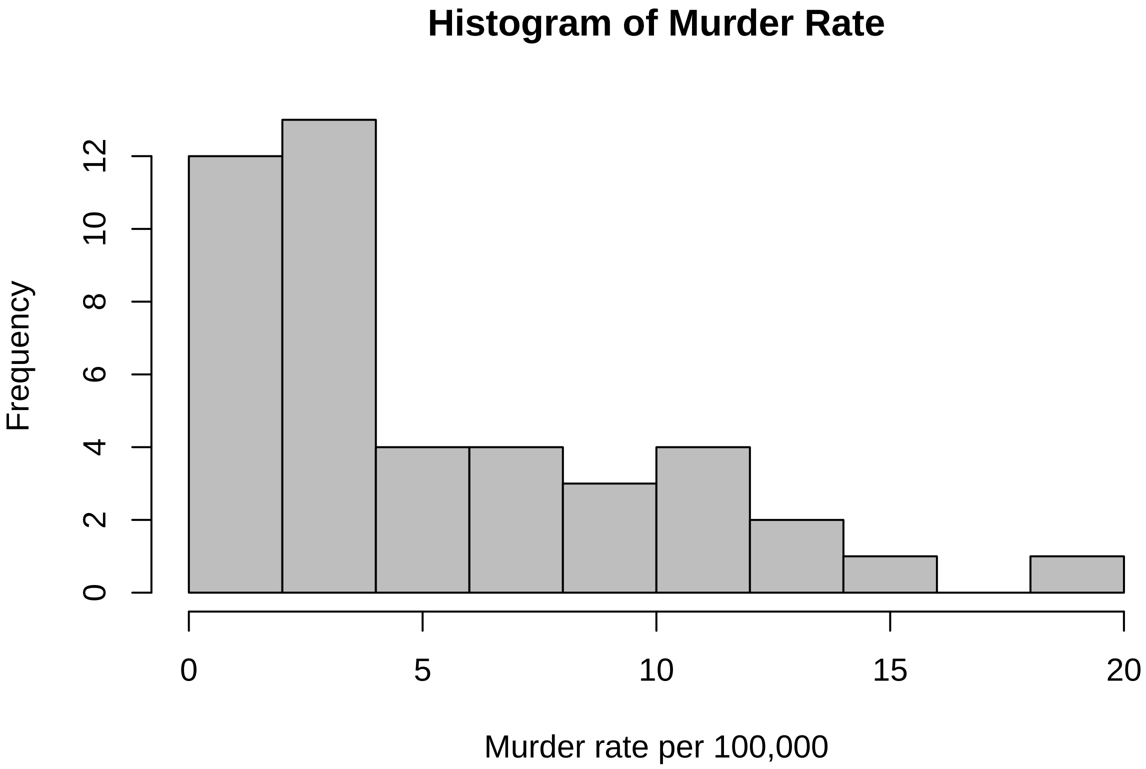
1.3 Descriptive Analysis of Variables

1.3.1 Graphs

The graph below illustrates a histogram displaying the frequencies with respect to the Murder Rate per 100,000. The histogram portrays the fact that most states have a murder rate of 0-4 per 100,000, with a mode of approximately between three and four.

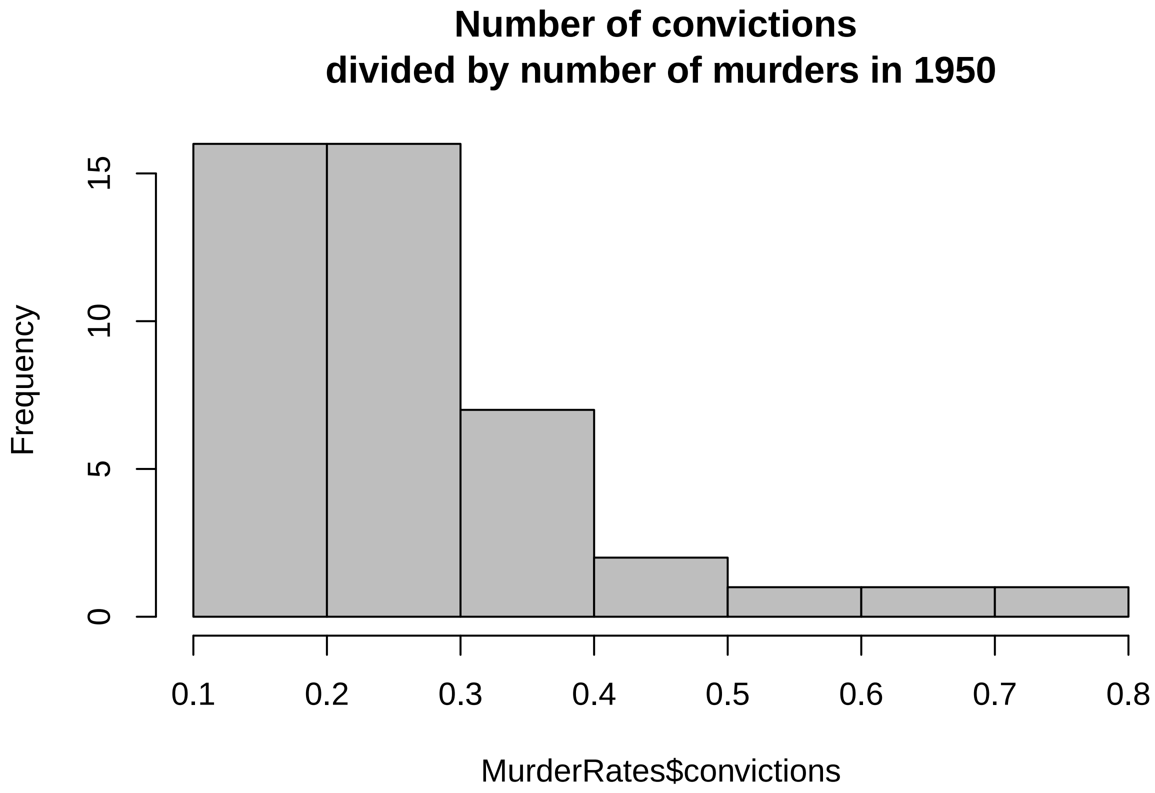
Notice that the histogram is right-skewed with a long tail towards the right. This resembles a specification error of a log normal distribution. We will correct for this specification error by taking the log of the log normal distribution to give us the normal distribution in the following sections.

```
#Histogram
hist(MurderRates$rate, col='grey',main = "Histogram of Murder Rate"
,xlab = "Murder rate per 100,000")
```

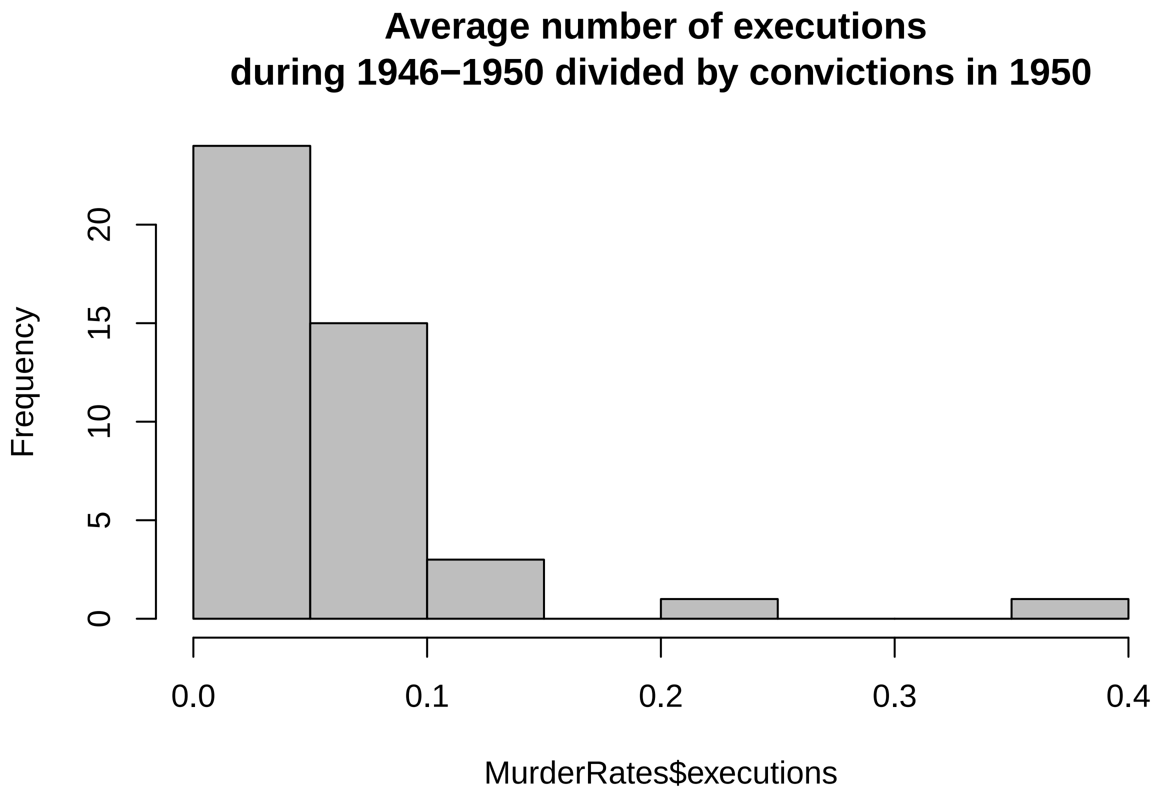


The other variables' histograms are presented below:

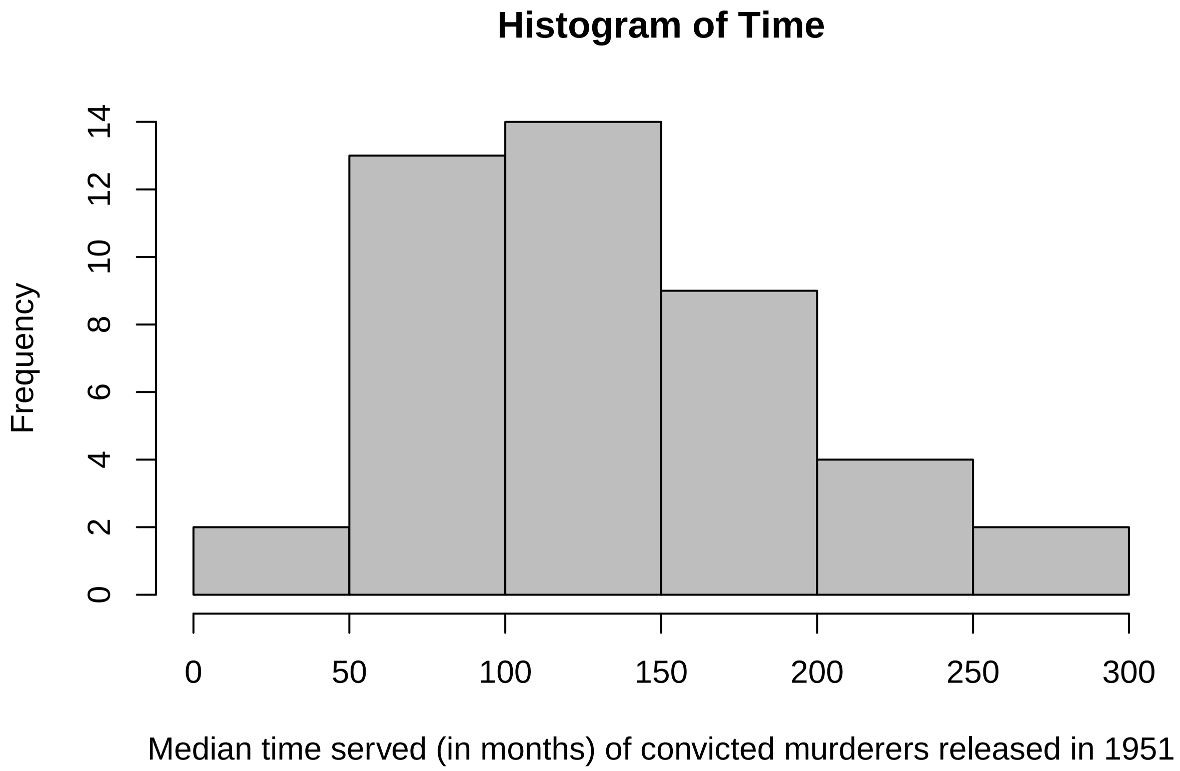
```
#Histogram
hist(MurderRates$convictions,
col='grey',main = "Number of convictions
divided by number of murders in 1950")
```



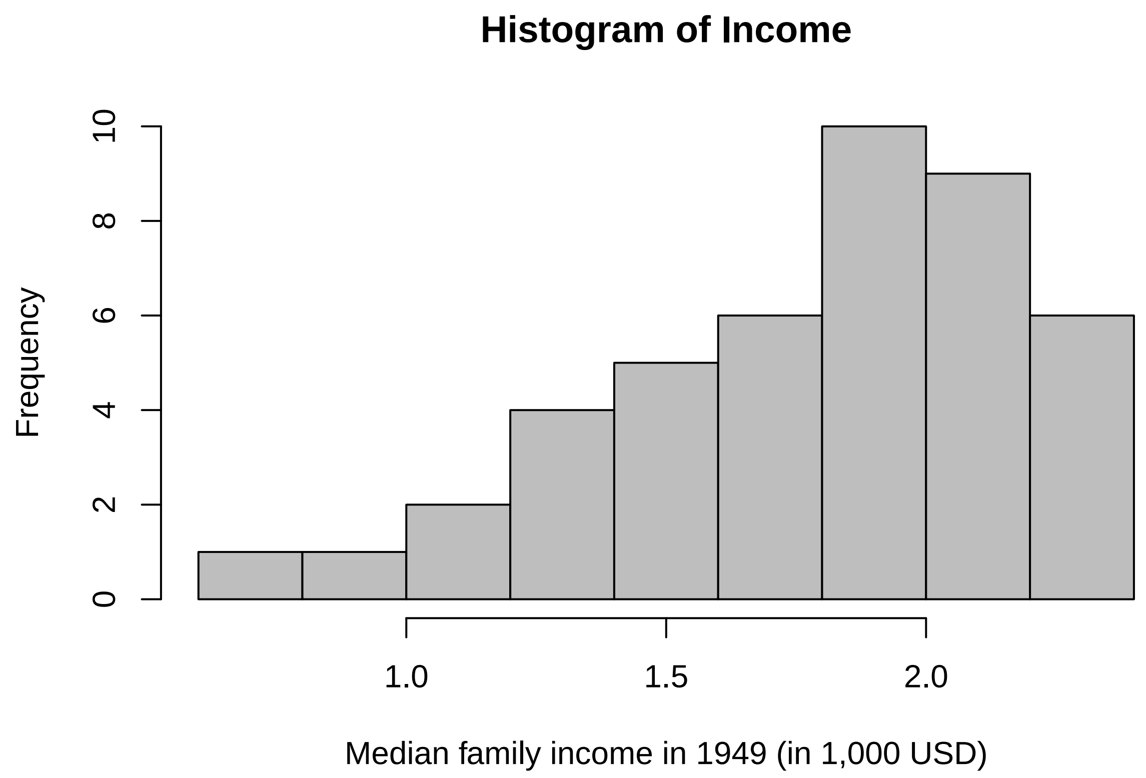
```r
#Histogram
hist(MurderRates$executions, col='grey',
main = "Average number of executions
during 1946-1950 divided by convictions in 1950")
```



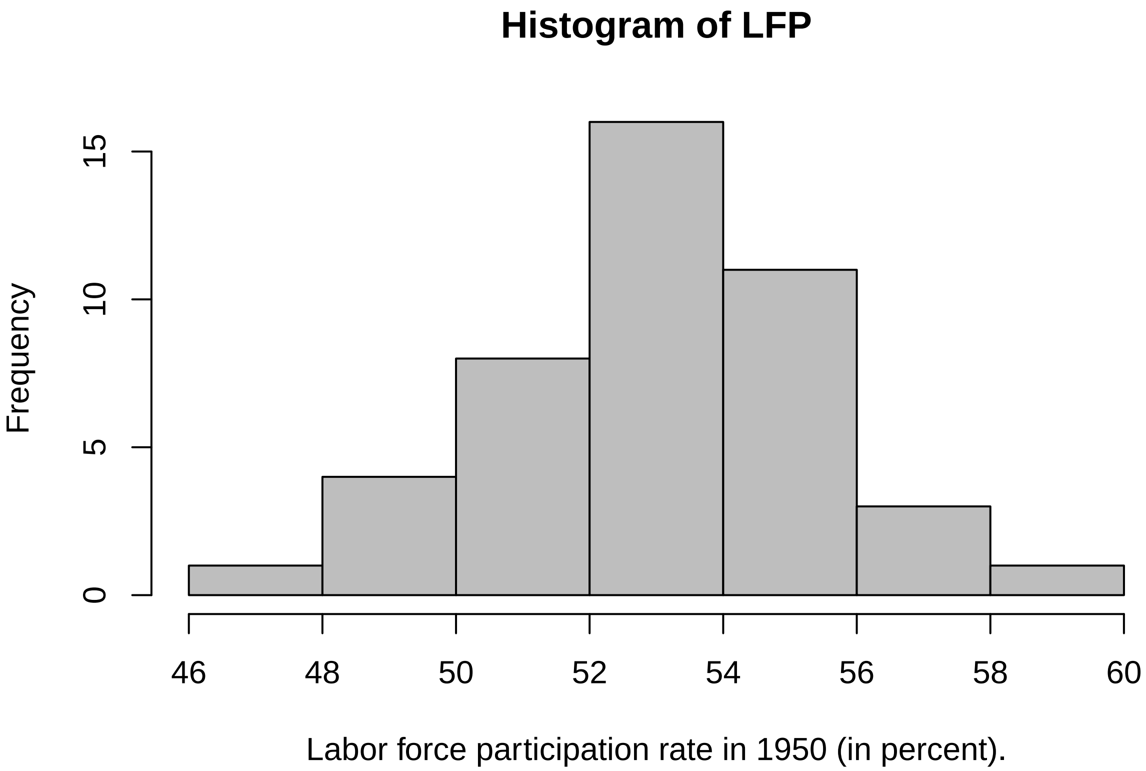
```
#Histogram
hist(MurderRates$time, col='grey',main = "Histogram of Time",
xlab = "Median time served (in months) of convicted murderers released in 1951")
```



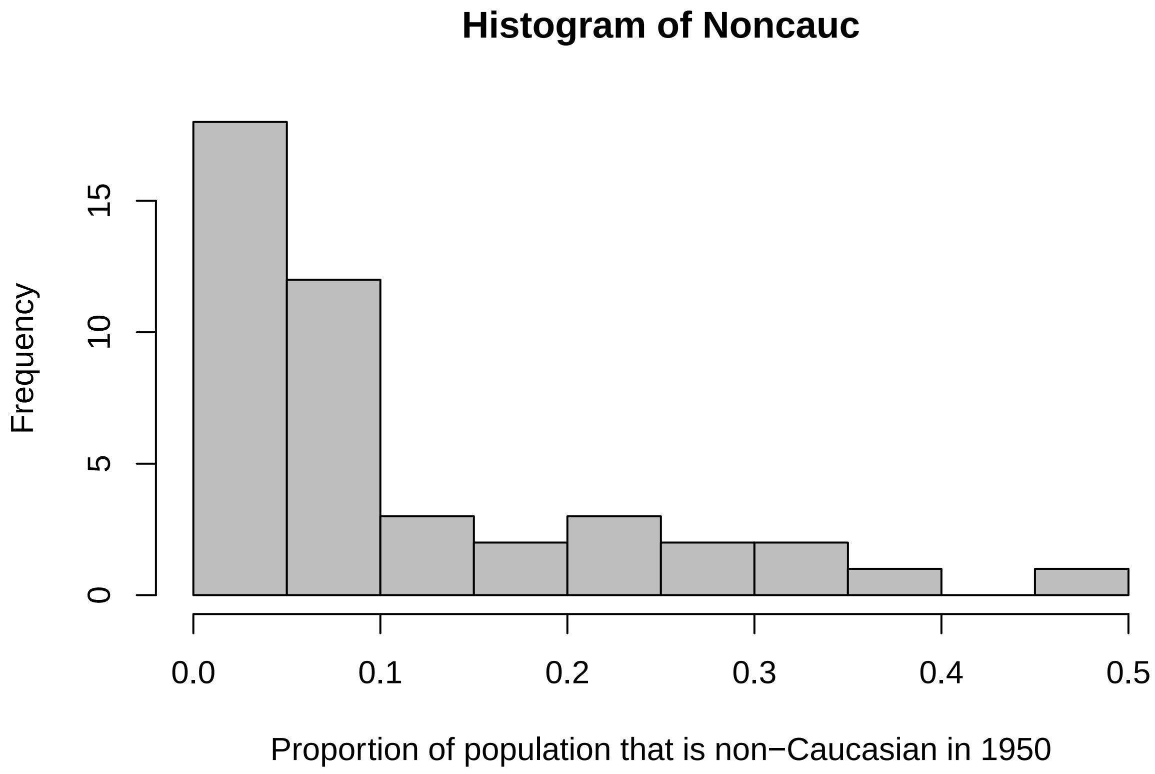
```r
#Histogram
hist(MurderRates$income, col='grey',main = "Histogram of Income",
xlab = "Median family income in 1949 (in 1,000 USD)")
```



```r
#Histogram
hist(MurderRates$lfp, col='grey',main = "Histogram of LFP",
xlab = "Labor force participation rate in 1950 (in percent).")
```



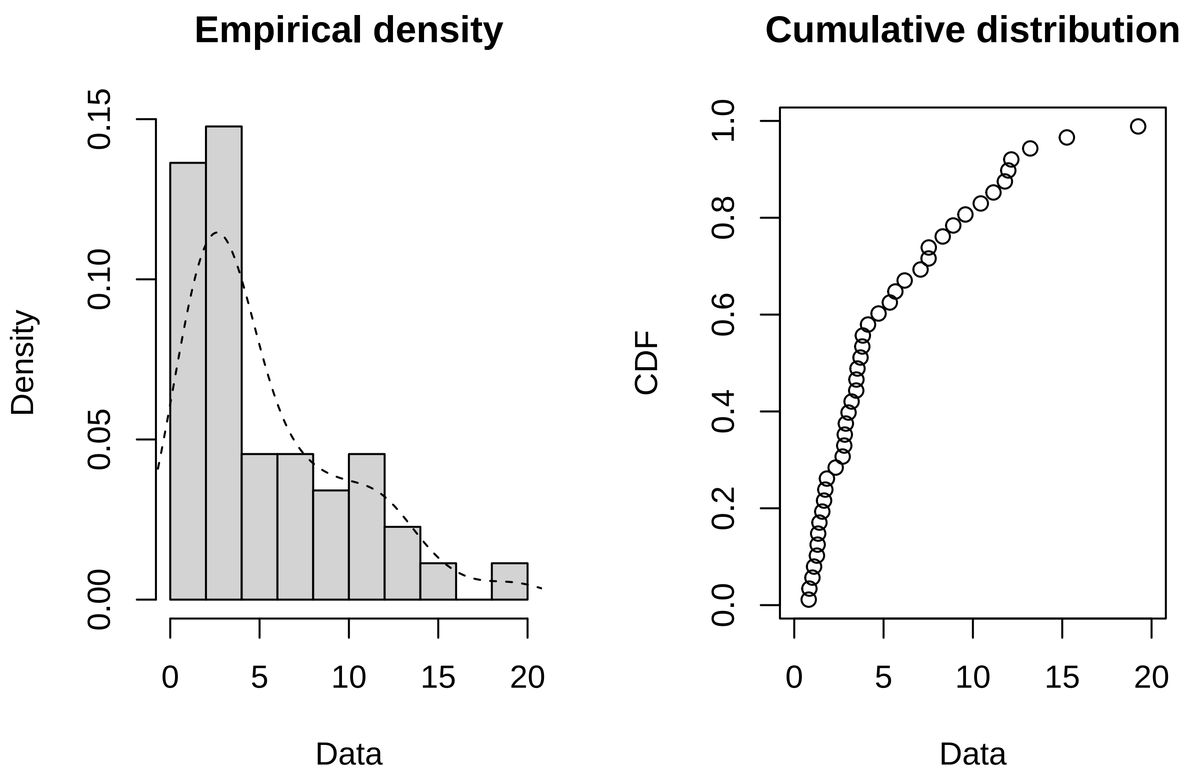
```
#Histogram
hist(MurderRates$noncauc, col='grey',main = "Histogram of Noncauc",
xlab = "Proportion of population that is non-Caucasian in 1950")
```



Delving deeper into the statistics, the graphs below shows the empirical density and cumulative distribution for the dependent variable. In particular, the fitted distributions support the previous fact that the dependent variable's central tendency is approximately centered around four because the cumulative distribution's 50th percentile is roughly around four.

Also, the fitted distributions also suggest that there is a specification error with the model because of the data being right-skewed.

```
#Fitted Distributions
plotdist(MurderRates$rate, histo = TRUE, demp = TRUE)
```



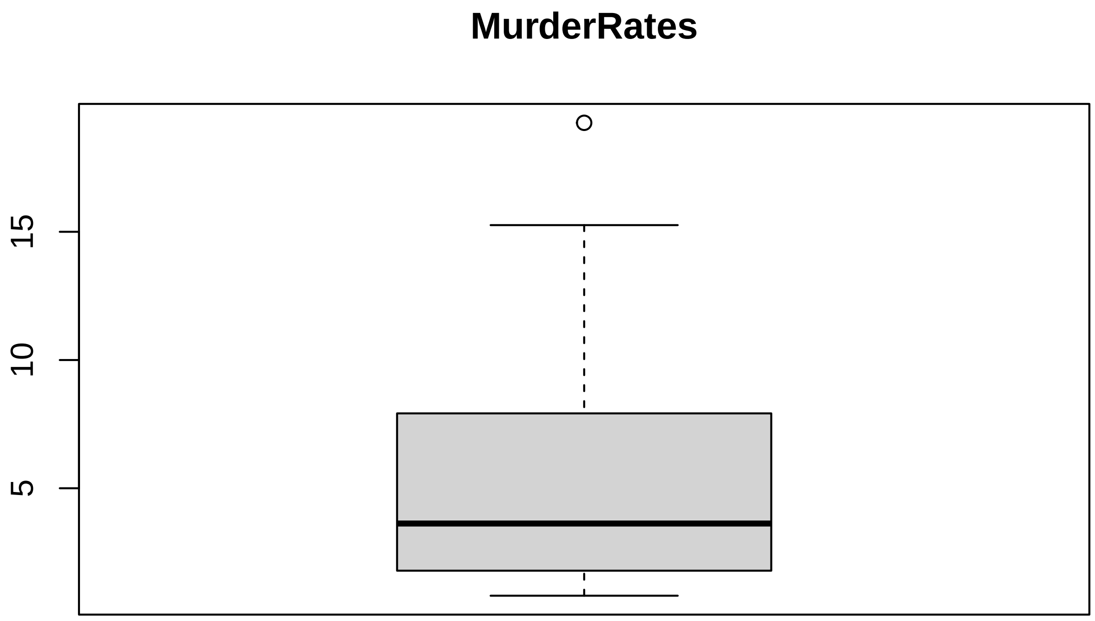
In order to create precise estimates about the statistics of the data set, we have included a five number summary below. The mean of the rate, convictions, executions, time, income, lfp, and noncauc variables are 5.404, 0.2605, 0.06034, 136.5, 1.7681, 53.07, and 0.10559, respectively. Also, the median of the rate, convictions, executions, time, income, lfp, and noncauc variables are 3.625, 0.2260, 0.045, 124, 1.83, 53.40, and 0.06450 respectively.

```r
rates_summary <- MurderRates[, c(1,2,3,4,5,6,7,8)]
summary(rates_summary)
```

```
##       rate         convictions       executions          time      
##  Min.   : 0.810   Min.   :0.1080   Min.   :0.00000   Min.   : 34.0  
##  1st Qu.: 1.808   1st Qu.:0.1663   1st Qu.:0.02625   1st Qu.: 94.0  
##  Median : 3.625   Median :0.2260   Median :0.04500   Median :124.0  
##  Mean   : 5.404   Mean   :0.2605   Mean   :0.06034   Mean   :136.5  
##  3rd Qu.: 7.725   3rd Qu.:0.3202   3rd Qu.:0.08225   3rd Qu.:179.0  
##  Max.   :19.250   Max.   :0.7570   Max.   :0.40000   Max.   :298.0  
##      income           lfp           noncauc        southern
##  Min.   :0.760   Min.   :47.00   Min.   :0.00300   no :29  
##  1st Qu.:1.550   1st Qu.:51.50   1st Qu.:0.02175   yes:15  
##  Median :1.830   Median :53.40   Median :0.06450           
##  Mean   :1.781   Mean   :53.07   Mean   :0.10559           
##  3rd Qu.:2.070   3rd Qu.:54.52   3rd Qu.:0.14450           
##  Max.   :2.390   Max.   :58.80   Max.   :0.45400           
```

Next, we present a Box Plot for the dependent variable. An interesting insight that the box plot provides us is that the data set contains a single outlier of 19.25 murders per 100,000.

```
#Box Plot For MurderRates
boxplot(MurderRates$rate, main="MurderRates")
```



Finally, we present the correlation matrix. The correlation between the murder rate and executions is 0.1727, a positive relationship. The correlation between the murder rate and convictions is -0.25113. The correlation between the murder rate and noncauc is 0.7486359. The correlation between the murder rate and time is -0.51858. The correlation between the murder rate and income is -0.65428. The correlation between the murder rate and lfp is -0.1827364.

```
#Correlation Matrix
my_data <- MurderRates[, c(1,2,3,4,5,6,7)]
cor(my_data)
```

```
##                   rate  convictions  executions         time      income
## rate         1.0000000 -0.251134671  0.17279303 -0.518584162 -0.65427652
## convictions -0.2511347  1.000000000 -0.21352746  0.004829982  0.06577502
## executions   0.1727930 -0.213527464  1.00000000  0.079556439  0.03783757
## time        -0.5185842  0.004829982  0.07955644  1.000000000  0.30356188
## income      -0.6542765  0.065775022  0.03783757  0.303561875  1.00000000
## lfp         -0.1827364 -0.172351045  0.29625365  0.154922595  0.55790866
## noncauc      0.7486359 -0.184331213  0.22077998 -0.335964017 -0.66062442
##                    lfp    noncauc
## rate        -0.1827364  0.7486359
## convictions -0.1723510 -0.1843312
## executions   0.2962537  0.2207800
## time          0.1549226 -0.3359640
## income        0.5579087 -0.6606244
## lfp           1.0000000 -0.1507521
## noncauc      -0.1507521  1.0000000
```

1.3.2 Possible Violations of Regression Assumptions

The most possible but apparent violation of the regression assumptions would be homoskedasticity. According to the linear regression assumption of homoskedasticity, linear regressions must have a constant variance in its error term. However, by analyzing the variables, we observe multiple clues towards heteroskedasticity.

First, the histogram reveals that the data set has an incorrect transformation of the dependent variable. The histogram is right-skewed with a long tail towards the right. This resembles a specification error of a log normal distribution and may indicate the presence of heteroskedasticity. We will need to correct for this specification error by taking the log of the log normal distribution to give us the normal distribution.

Second, the box plot reveals that there is an outlier in the dependent variable. Although the 3rd quartile of the rates variable is 7.725 per 100,000, the data set contains a data point of 19.25 murders per 100,000. An outlier may cause heteroskedasticity.

Third, the mixing of observation of different scales in the income variable may cause heteroskedasticity. The income variable ranges from 0.760 to 2.390 (in 1,000 USD). The mixing of high-income households with low-income households may cause heteroskedasticity.

---

2 The Model

2.1 The Multiple Linear Regression Model

2.1.1 Model and Inference

We will estimate the relationship between the rate, execution, time, income, convictions with the following Multiple Regression Model:

$$MURDERRATES = \beta_1 + \beta_2 INCOME + \beta_3 TIME + \beta_4 EXECUTION + \beta_5 CONVICTIONS + \beta_6 LFP + \beta_7 NONCAUC + \beta_8 SOUTHERN + \epsilon_i$$

Our claim about the income variable having an equal effect in decreasing rates with the time variable can be precisely modeled by the following inference:

$$H_0 : \beta_2 = \beta_3 H_1 : \beta_2 \neq \beta_3$$

```r
library(AER)  # Load the AER package
data("MurderRates")
```

xIn this following model: "rate" is the response variable (Murder Rates), which is the variable I want to predict. "convictions," "executions," "times," "lfp," and "income" are the predictor variables I want to include in the model. These variables will be used to explain or predict the variation in house prices.

```
str(MurderRates)
```

```
## ’data.frame’:    44 obs. of  8 variables:
##  $ rate       : num  19.25 7.53 5.66 3.21 2.8 ...
##  $ convictions: num  0.204 0.327 0.401 0.318 0.35 0.283 0.204 0.232 0.199 0.138 ...
##  $ executions : num  0.035 0.081 0.012 0.07 0.062 0.1 0.05 0.054 0.086 0 ...
##  $ time       : int  47 58 82 100 222 164 161 70 219 81 ...
##  $ income     : num  1.1 0.92 1.72 2.18 1.75 2.26 2.07 1.43 1.92 1.82 ...
##  $ lfp        : num  51.2 48.5 50.8 54.4 52.4 56.7 54.6 52.7 52.3 53 ...
##  $ noncauc    : num  0.321 0.224 0.127 0.063 0.021 0.027 0.139 0.218 0.008 0.012 ...
##  $ southern   : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 2 1 1 ...
```

```
Reg <- lm(rate~income+time+executions+convictions+lfp+
noncauc+southern, data = MurderRates)
summary(Reg)
```

```
## 
## Call:
## lm(formula = rate ~ income + time + executions + convictions + 
##     lfp + noncauc + southern, data = MurderRates)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.9913 -1.1943 -0.3538  1.2383  6.5574 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  0.44436    9.96694   0.045   0.9647  
## income      -2.50013    1.68519  -1.484   0.1466  
## time        -0.01547    0.00705  -2.194   0.0348 *
## executions   2.85276    6.12313   0.466   0.6441  
## convictions -4.33938    2.78313  -1.559   0.1277  
## lfp          0.19357    0.20614   0.939   0.3540  
## noncauc     10.39903    5.40610   1.924   0.0623 .
## southernyes  3.26216    1.32980   2.453   0.0191 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 2.459 on 36 degrees of freedom
## Multiple R-squared:  0.7459, Adjusted R-squared:  0.6965 
## F-statistic:  15.1 on 7 and 36 DF,  p-value: 5.105e-09
```

Commenting on the overall fit of the model:

R-Squared : The model has an R-squared which is 0.6327. This is a reasonably good R-squared value. Adjusted R-squared is 0.5844 providing slightly more conservative of model fit.

F-statistic: F-statistic is 13.09 and a very low p-value is 1.953e-07 which indicates that the model is statistically significant since p-value is lower than the F-statistic.

Coefficient significance: Coefficients provide information about their statistical significance. From the model, income and time have highly significant coefficients with very low p-values. For that reason, if income or time changes, others predictor will have effect.

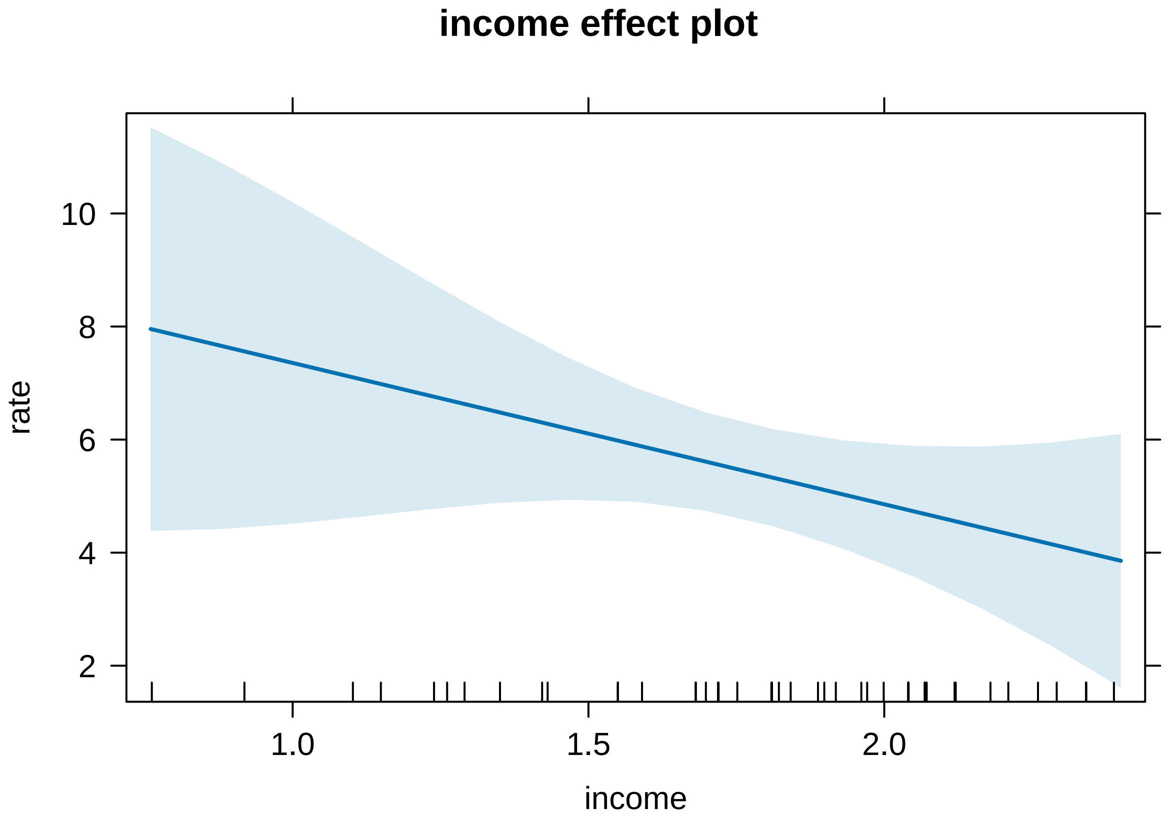
Based on the summary regression above, the modeled relationship between the rate, execution, time, income, convictions becomes:

$$MURDERRATES = 7.650 + (-7.038)INCOME + (-0.026)TIME + 9.634EXECUTION + (-4.739)CONVICTIONS + 0.274LFP + (10.399)NONCAUC + (3.262)SOUTHERN + \epsilon_i$$

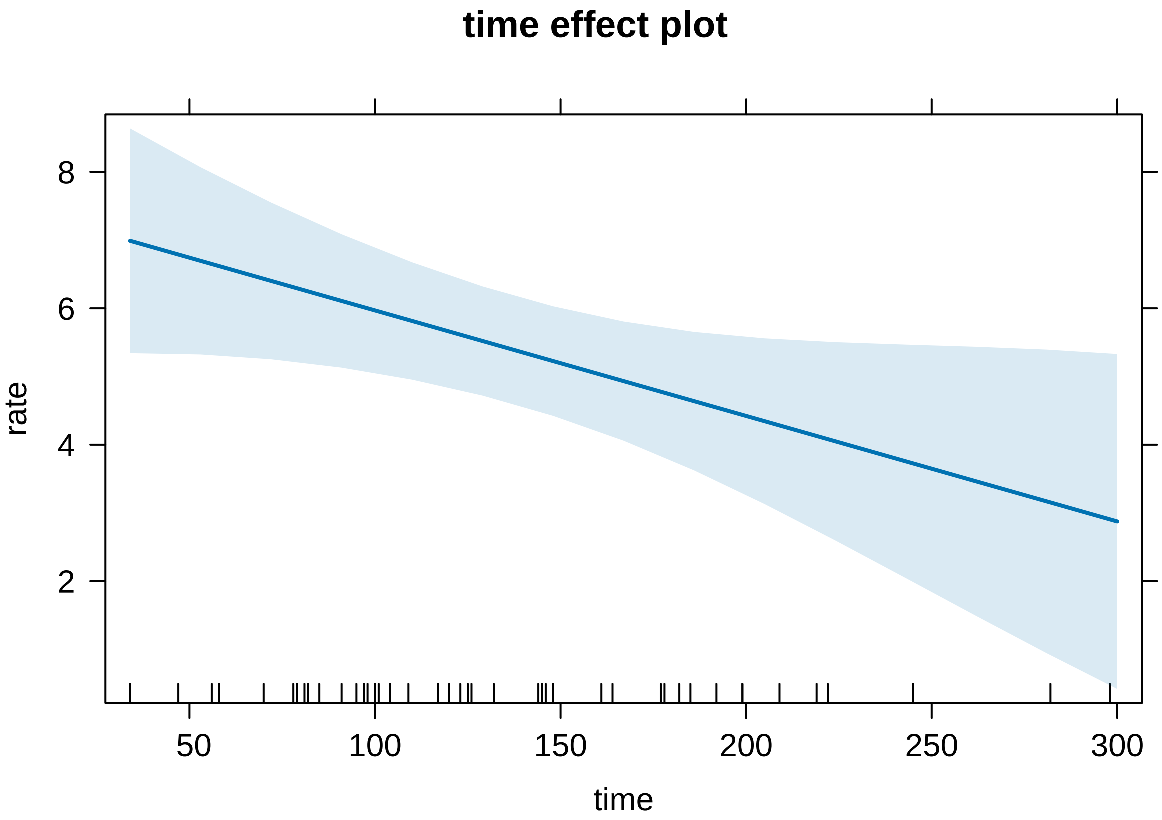
An unexpected finding from the data is that the coefficient for the execution variable is positive, indicating a positive relationship between Murder Rates and Executions. This is surprising because the purpose of death penalty executions are to prevent future crime. Yet, our data indicators otherwise.

The following are the partial effects of the independent variables on the rates variable. This will help us assess the marginal change of each individual independent variable with its respect to the rates variable.

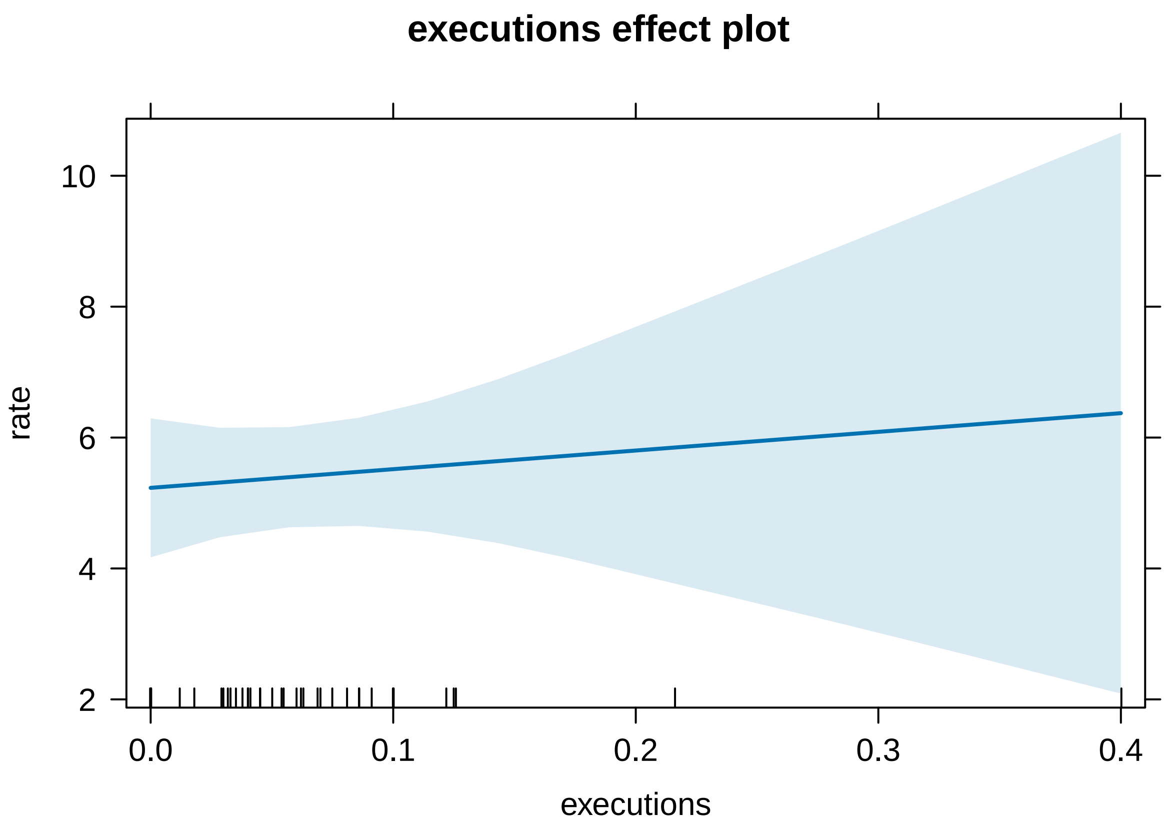
```r
#Partial effect of Income on Rates
#install.packages("effects")
library(effects)
effincome <- effect("income", Reg)
plot(effincome)
```



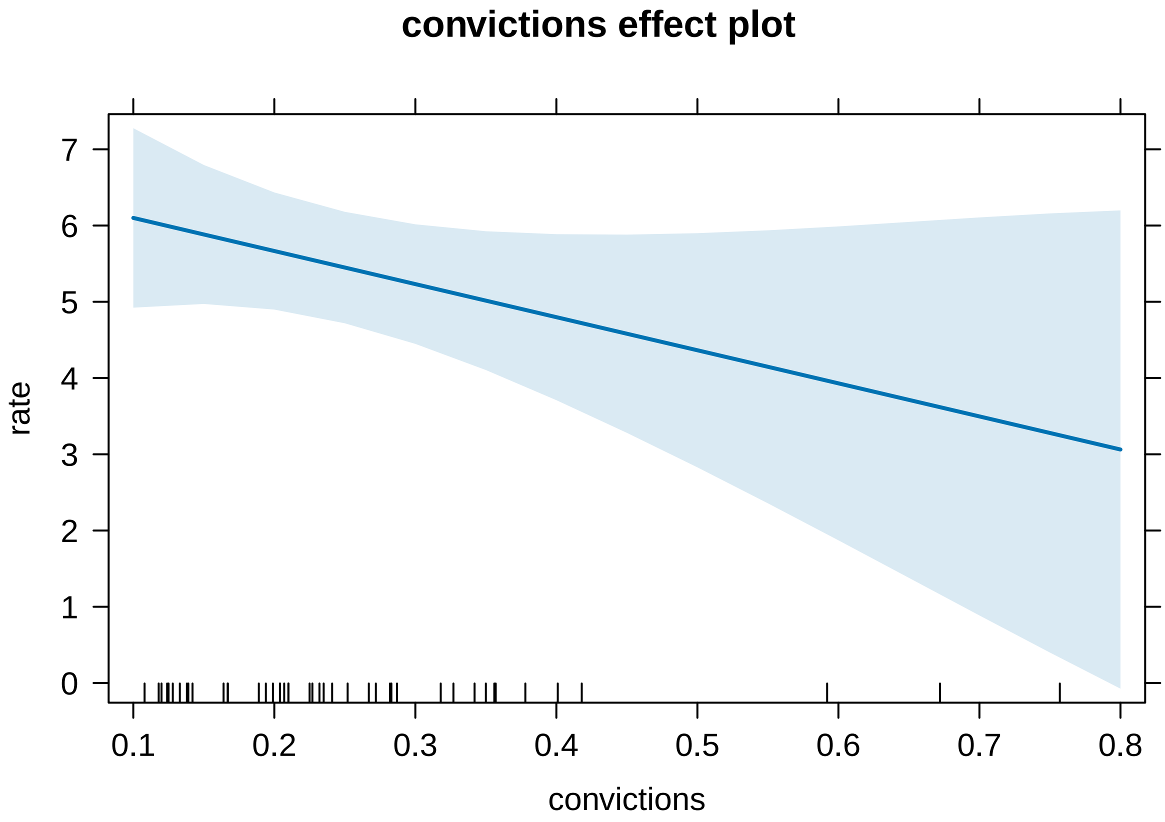
```r
#Partial effect of Executions on Rates
efftime <- effect("time", Reg)
plot(efftime)
```



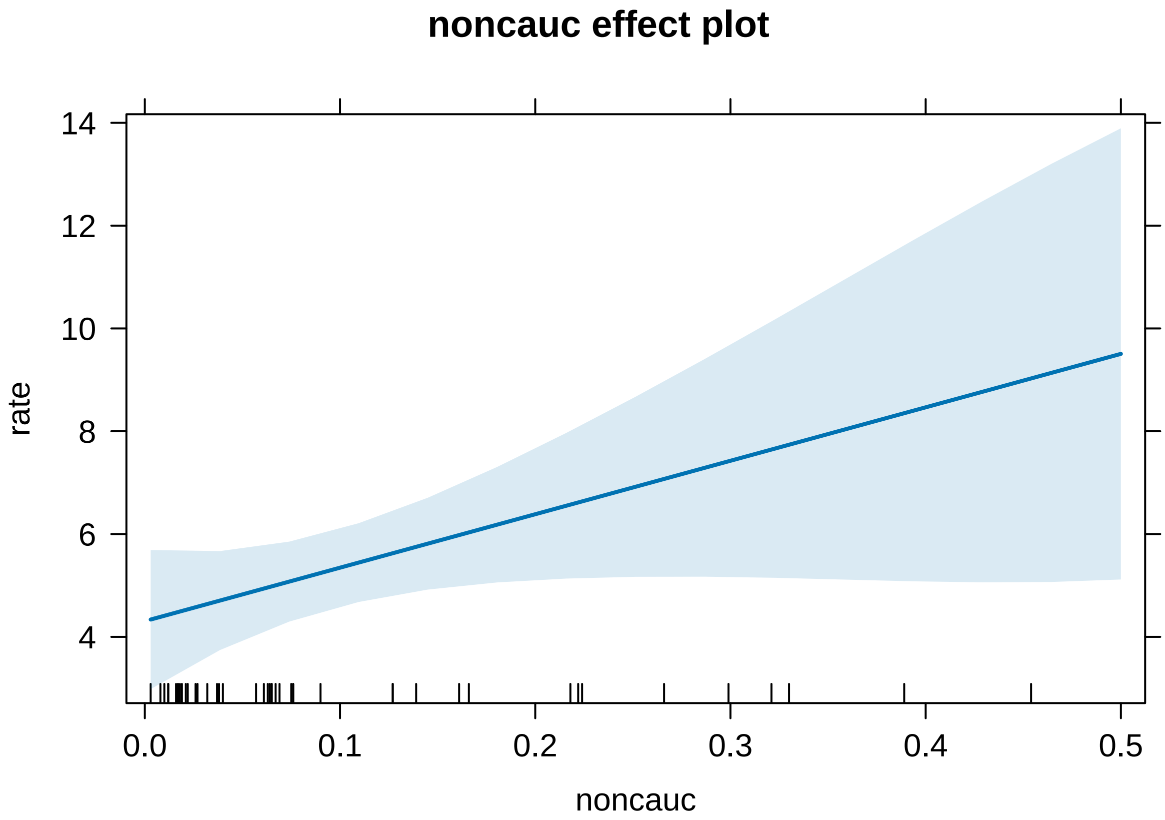
```
#Partial effect of Executions on Rates
effexecutions <- effect("executions", Reg)
plot(effexecutions)
```



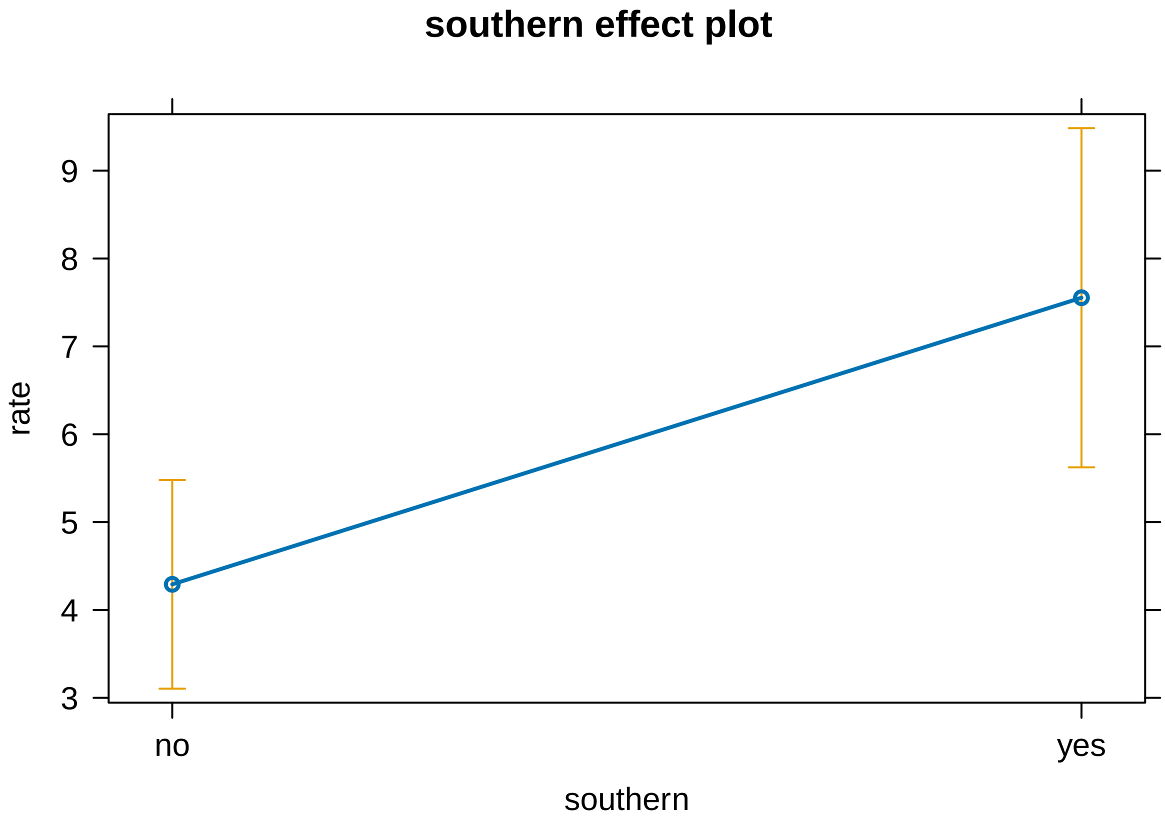
```
#Partial effect of Executions on Rates
effconvictions <- effect("convictions", Reg)
plot(effconvictions)
```



```
#Partial effect of Executions on Rates
effnoncauc <- effect("noncauc", Reg)
plot(effnoncauc)
```



```
#Partial effect of Executions on Rates
effsouthern <- effect("southern", Reg)
plot(effsouthern)
```



2.1.2 Testing Multicollinearity

The following is a summary for Multicollinearity using VIF:

```
library(car)
vif_values <- vif(Reg)  # Assuming 'Reg' is your original regression model
summary(vif_values)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.106   1.299   1.859   2.046   2.796   3.168
```

Based on the VIF analysis, I don't need to remove any variables from my model since VIF values are within range. ##### Here is why: Min: The minimum VIF is 1.096 which means low multicollinearity.

1st Qu.: 1st 25 percentile of all VIF values is 1.109 which is close to 1. That's mean this is also under low multicollinearity.

Median: 1.157 is the median of all variables which means median is also under multicollinearity.

Mean: Mean with 1.345 indicates relatively low of multicollineairity.

3rd Qu.: The VIF value at the 75th percentile of your predictor variables is approximately 1.660, which is still relatively close to 1. This indicates that the majority of your predictor variables have low multicollinearity.

Max: The maximum value among all the predictors is 1.700 which is not present highly multicollinearity. This indicates that all predictor's VIF is under 1.7 which is low multicollinearity.

Because of this result, we don't need to remove any variables.

Min. 1st Qu. Median Mean 3rd Qu. Max. 1.096 1.109 1.157 1.345 1.660 1.700

. . . . .

2.2 The New Model

2.2.1 Akaike Information Criterion

To create a new model, we need to observe other regression models with packege.

```
#Model1 <- lm(rate~income+time+executions+convictions+lfp
#+noncauc+southern, data=MurderRates)

#aic_value <- AIC(Model1)
#aic_value

AIC1 <- lm(rate~income, data=MurderRates)
AIC2 <- lm(rate~income+time, data=MurderRates)
AIC3 <- lm(rate~income+time+executions,data=MurderRates)
AIC4 <- lm(rate~income+time+executions+convictions+lfp, data=MurderRates)
AIC5 <- lm(rate~income+time+executions+convictions+lfp+noncauc,data=MurderRates)
AIC6 <- lm(rate~income+time+executions+convictions+lfp+noncauc+southern,data=MurderRates)

value1 <- AIC(AIC1)
value2 <- AIC(AIC2)
value3 <- AIC(AIC3)
value4 <- AIC(AIC4)
value5 <- AIC(AIC5)
value6 <- AIC(AIC6)

value1
```

```
## [1] 236.9114
```

```
value2
```

```
## [1] 229.2478
```

```
value3
```

```
## [1] 226.2389
```

```
value4
```

```
## [1] 225.4212
```

```
value5
```

```
## [1] 218.0164
```

```
value6
```

```
## [1] 213.215
```

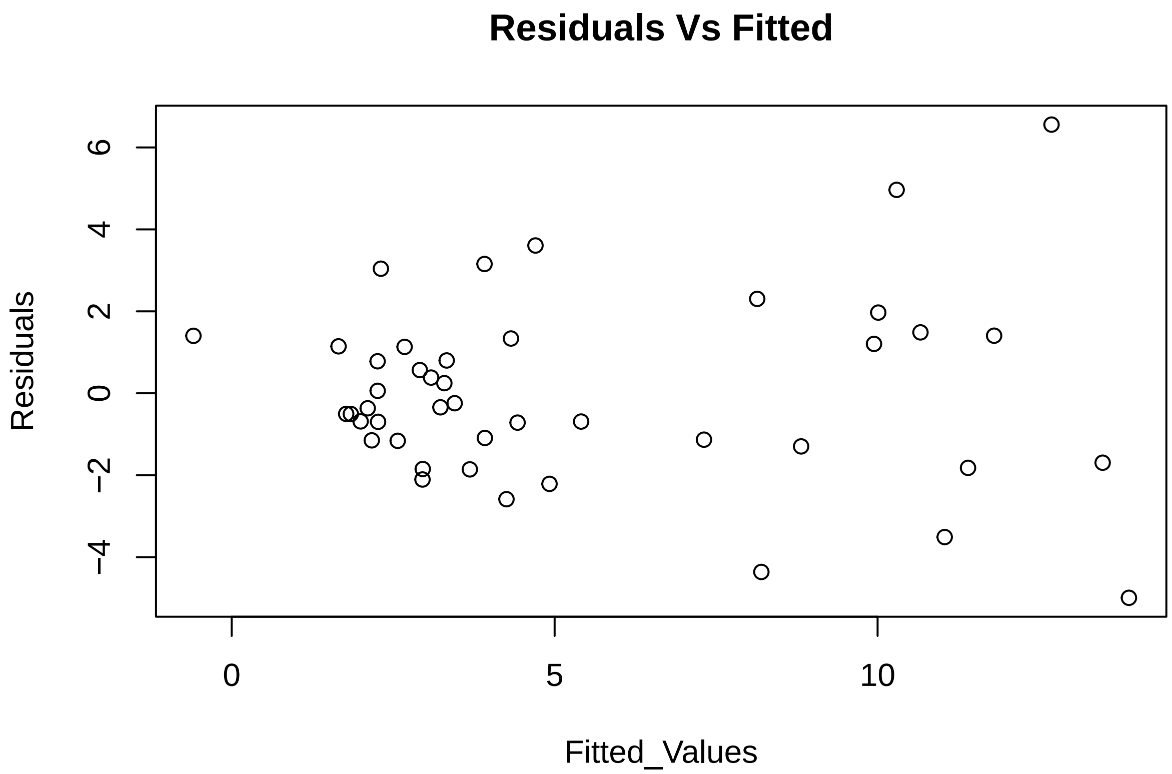
As we observe, the AIC was obtained by adding the variables from income to southern. As a result, the AIC decreased with each addition of the variables. This is a good result for AIC, where lower values are good-fit. Therefore, we can conclude that the regression model we used is correct.

2.2.2 Plotting Residuals Versus Fitted Values

```
Model1 <- lm(rate~income+time+executions+convictions+lfp+noncauc+southern,data=MurderRates)

residuals <-residuals (Model1)
fitted_values <- fitted(Model1)

plot(fitted_values, residuals, main="Residuals Vs Fitted",
xlab="Fitted_Values", ylab="Residuals" )
```



From this plot, as the value of x increases, there is a variation in the value of y. That means there is a Heteroskedasticity .

2.2.3 RESET Test

Marc: Step 7

We will proceed to run a RESET test to the second power on our regression model in order to test whether or not the mode has a wrong functional form. The original and modified model is denoted below:

$$MURDERRATES = \beta_1 + \beta_2 INCOME + \beta_3 TIME + \beta_4 EXECUTION + \beta_5 CONVICTIONS + \beta_6 LFP + \beta_7 NONCAUC + \beta_8 SOUTHERN + \beta_9 \hat{y}^2 + \epsilon_i$$

The RESET Hypothesis Test is as follows:

$$H_0 : \beta_9 = 0 H_1 : \beta_9 \neq 0$$

Running the RESET test as indicated below at a 5% significance level, we obtain the following statistics:

```
resettest(Reg, power=2, type="regressor", data=MurderRates)
```

```
##
##  RESET test
##
## data:  Reg
## RESET = 2.9634, df1 = 6, df2 = 30, p-value = 0.02156
```

Since the p-value of the model is 0.02156, it is below the 5% significance level that we assumed when running the RESET test. Therefore, we will need to reject the null hypothesis that the coefficient of the second power term is 0. This tells us that our model is in an incorrect functional form with the model being linear according to the RESET test. Thus, we would need to add higher order terms in our model. For instance, adding the higher order term for the income variable in the following regression would increase the p-value to 0.08579, which is above our significance level of 5%.

```
New_Reg <- lm(rate~income+time+executions+convictions+
lfp+noncauc+southern+poly(income,2), data = MurderRates)
resettest(New_Reg, power=2, type="regressor", data=MurderRates)
```

```
##
##  RESET test
##
## data:  New_Reg
## RESET = 2.0073, df1 = 8, df2 = 26, p-value = 0.08579
```

---

3 Heteroskedasticity

3.1 Testing For Heteroskedasticity

We are using the Breusch-Pagan test to see if our model has heteroskedasticity. Based on the residual plot above, we suspect that there is some heteroskedasticity. When the test is performed, we get a p-value of 0.03152. If we compare this to a 5% significance level, we reject the null, and we can conclude that this model has heteroskedasticity.

```r
#install.packages("lmtest")
library(lmtest)

Reg <- lm(rate~income+time+executions+convictions+
lfp+noncauc+southern,data=MurderRates)

# Breusch-Pagan test
bp_test1 <- bptest(Reg)

# results of Breusch-Pagan test
print(bp_test1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Reg
## BP = 15.372, df = 7, p-value = 0.03152
```

To correct the heteroskedasticity, we can use the feasible GLS method. We could have used the Weighted LS method, but we were unsure of the form of the skedastic function. Hence, the feasible GLS method is a more general way to correct the model. After performing the feasible GLS method, we can again perform a Breusch-Pagan test. This time, we get a p-value of 0.9534. Comparing this to a significance level of 5%, we can fail to reject the null and conclude that we do not have heteroskedasticity.

```r
# Feasible GLS
ehatsq <- resid(Reg)^2
sighatsq.ols  <- lm(log(ehatsq)~income+time+executions+convictions
+lfp+noncauc+southern,data=MurderRates)
vari <- exp(fitted(sighatsq.ols))
Reg.fgls <- lm(rate ~ income+time+executions+
convictions+lfp+noncauc+southern, weights=1/vari,data=MurderRates)
summary(Reg.fgls)
```

```
##
## Call:
## lm(formula = rate ~ income + time + executions + convictions +
##     lfp + noncauc + southern, data = MurderRates, weights = 1/vari)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4070 -1.1935 -0.0332  0.9731  4.8150
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.641069   6.800758   0.241 0.810687
## income      -2.802950   1.091940  -2.567 0.014562 *
## time        -0.008124   0.004076  -1.993 0.053875 .
## executions   2.787635   5.376568   0.518 0.607297
## convictions -4.400798   1.628355  -2.703 0.010430 *
## lfp          0.161847   0.138007   1.173 0.248599
## noncauc     11.213849   3.038471   3.691 0.000736 ***
## southernyes  2.802642   0.901766   3.108 0.003668 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.638 on 36 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6017
## F-statistic: 10.28 on 7 and 36 DF,  p-value: 5.243e-07
```

```r
bptest(Reg.fgls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Reg.fgls
## BP = 2.1125, df = 7, p-value = 0.9534
```

. . . . .

3.2 Correcting For Heteroskedasticity

We can select a model using backward selection. This model (Reg4) led to a great adjusted R-squared of 0.9179. The p-value on the Breusch-Pagan test was 0.4627, which, when compared to a significance level of 5%, we can fail to reject the null and conclude that we do not have heteroskedasticity. Based on the previous discussion of the data and the fact that it is skewed right, we wanted to test the log-normal model. This model (Reg6) had an adjusted R-squared of 0.7725. The p-value was 0.4538, which, when compared to a significance level of 5%, we can fail to reject the null and conclude that we do not have heteroskedasticity. The AIC and BIC in the first regression were greater than the second regression, 161.8709 and 61.11577, respectively, for AIC, and 199.3389 and 93.23118, respectively, for BIC. This could mean that the first model was overfitting the data because it had more terms, but the AIC/BIC values penalize having those extra terms. Thus, the model in Reg6 is a better model.

```r
Reg3 <- lm(rate~income+time+executions+convictions+lfp+noncauc+southern
+(income+time+executions+convictions
+lfp+noncauc+southern)^2, data=MurderRates)
Reg4 <- stepAIC(Reg3, direction = "backward", trace = FALSE)

summary(Reg4)
```

```
##
## Call:
## lm(formula = rate ~ income + time + executions + convictions +
##     lfp + noncauc + southern + income:time + income:executions +
##     income:convictions + income:noncauc + time:executions + time:convictions +
##     executions:convictions + executions:southern + convictions:lfp +
##     convictions:noncauc + convictions:southern + noncauc:southern,
##     data = MurderRates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8689 -0.5629 -0.1373  0.4820  2.3624
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -33.82085   19.26602  -1.755 0.091937 .
## income                   -22.95371    4.44817  -5.160 2.77e-05 ***
## time                      -0.10089    0.02900  -3.479 0.001939 **
## executions               202.85273   72.58413   2.795 0.010052 *
## convictions              153.75095   56.44443   2.724 0.011838 *
## lfp                        1.51601    0.35887   4.224 0.000298 ***
## noncauc                  -38.32751   23.72004  -1.616 0.119202
## southernyes               11.46278    3.07316   3.730 0.001039 **
## income:time                0.05052    0.01491   3.388 0.002427 **
## income:executions        -43.41065   25.50053  -1.702 0.101611
## income:convictions        43.38518   10.94560   3.964 0.000577 ***
## income:noncauc            21.68673    9.71722   2.232 0.035227 *
## time:executions           -0.33458    0.12255  -2.730 0.011672 *
## time:convictions           0.04024    0.02496   1.612 0.119955
## executions:convictions  -200.21788   79.74346  -2.511 0.019188 *
## executions:southernyes  -164.91129   33.22915  -4.963 4.57e-05 ***
## convictions:lfp           -4.59750    1.16516  -3.946 0.000604 ***
## convictions:noncauc       59.22730   44.39598   1.334 0.194703
## convictions:southernyes  -20.70680    6.75535  -3.065 0.005312 **
## noncauc:southernyes       17.32971   12.24006   1.416 0.169676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.279 on 24 degrees of freedom
## Multiple R-squared:  0.9542, Adjusted R-squared:  0.9179
## F-statistic: 26.29 on 19 and 24 DF,  p-value: 1.159e-11
```

```
bptest(Reg4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Reg4
## BP = 18.908, df = 19, p-value = 0.4627
```

```
Reg5 <- lm(log(rate)~income+time+executions+convictions+lfp+noncauc+southern
           +(income+time+executions+convictions+lfp+noncauc+southern)^2, data=MurderRates)
Reg6 <- stepAIC(Reg5, direction = "backward", trace = FALSE)

summary(Reg6)
```

```
##
## Call:
## lm(formula = log(rate) ~ income + time + executions + convictions +
##     lfp + noncauc + southern + income:executions + income:convictions +
##     income:noncauc + time:executions + executions:convictions +
##     executions:southern + convictions:lfp + convictions:noncauc +
##     convictions:southern, data = MurderRates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71848 -0.10966 -0.03364  0.15563  0.74134
##
```

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -8.666e+00  5.542e+00  -1.564  0.12954
## income                  -1.264e+00  9.564e-01  -1.321  0.19743
## time                    -2.079e-04  1.979e-03  -0.105  0.91713
## executions               5.025e+01  2.259e+01   2.224  0.03467 *
## convictions              3.362e+01  1.772e+01   1.898  0.06850 .
## lfp                      2.281e-01  1.084e-01   2.104  0.04479 *
## noncauc                 -6.301e+00  4.650e+00  -1.355  0.18668
## southernyes              2.475e+00  8.104e-01   3.053  0.00504 **
## income:executions       -1.213e+01  7.748e+00  -1.566  0.12906
## income:convictions       3.957e+00  3.100e+00   1.276  0.21265
## income:noncauc           3.511e+00  1.886e+00   1.862  0.07359 .
## time:executions         -8.646e-02  3.888e-02  -2.224  0.03472 *
## executions:convictions  -3.525e+01  2.518e+01  -1.400  0.17286
## executions:southernyes  -2.828e+01  1.016e+01  -2.782  0.00973 **
## convictions:lfp         -8.102e-01  3.516e-01  -2.304  0.02914 *
## convictions:noncauc      2.086e+01  1.315e+01   1.587  0.12426
## convictions:southernyes -3.974e+00  2.109e+00  -1.884  0.07040 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4109 on 27 degrees of freedom
## Multiple R-squared:  0.8571, Adjusted R-squared:  0.7725
## F-statistic: 10.13 on 16 and 27 DF,  p-value: 1.37e-07
```

```
#BP Test
bptest(Reg6)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Reg6
## BP = 15.989, df = 16, p-value = 0.4538
```

```
#AIC Reg4
AIC(Reg4)
```

```
## [1] 161.8709
```

```
#AIC Reg6
AIC(Reg6)
```

```
## [1] 61.11577
```

```
#BIC Reg4
BIC(Reg4)
```

```
## [1] 199.3389
```

```r
#BIC Reg6
BIC(Reg6)
```

```
## [1] 93.23118
```

---

4 Conclusion

Finally, going back to our initial question, does the income variable have an equal effect with the time variable in decreasing murder rates? More precisely, we will conduct the hypothesis test below on the "Reg6" model we found after correcting for heteroskedasticity in Section 3.2. The hypothesis test is as follows:

$$H_0 : \beta_2 = \beta_3 H_1 : \beta_2 \neq \beta_3$$

Since we are testing hypotheses for multiple variables, we will use the F-test as a test statistic to determine whether to accept or reject the null hypothesis. We proceed to perform the hypothesis test below:

```r
hypothesis <- "income = time" # Using an equivalent equation
(test <- linearHypothesis (Reg6, hypothesis))
```

```
## Linear hypothesis test
##
## Hypothesis:
## income - time = 0
##
## Model 1: restricted model
## Model 2: log(rate) ~ income + time + executions + convictions + lfp +
##     noncauc + southern + income:executions + income:convictions +
##     income:noncauc + time:executions + executions:convictions +
##     executions:southern + convictions:lfp + convictions:noncauc +
##     convictions:southern
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 4.8536
## 2     27 4.5591  1   0.29453 1.7443 0.1977
```

```r
kable(test, caption="The `linearHypothesis()` object")
```

Table 1: The `linearHypothesis()` object

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
28	4.853631	NA	NA	NA	NA
27	4.559101	1	0.2945295	1.744268	0.1976846

Since the p-value of the two-tail linear hypothesis test at 5% significance level is 0.1949, the data indicates that we accept the null hypothesis. Thus, our initial hypothesis stated in the introduction is correct, and the income variable will have an equal effect with the time variable in decreasing murder rates.