Project 1

Rachel Asher, Caden Campbell, Juliesen Jaime

2023-10-12

1. Briefly discuss the question you are trying to answer with your model.

Our group is interested in knowing if income has an affect on the number of convictions for murder.

Question: Does income influence murder convictions?

```
##       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
```

2. Give a description of your dataset including:

(a) Citing the dataset

(b) A summary of what the dataset is about

(c) Descriptive analysis of your variables. This should include histograms with fitted distributions and correlation matrix, and the five number summary (which can be accompanied by a boxplot). All figures must include comments including, but not limited to, the distribution, central tendency and dispersion of the variables.

(d) Possible violation of the regression assumptions.

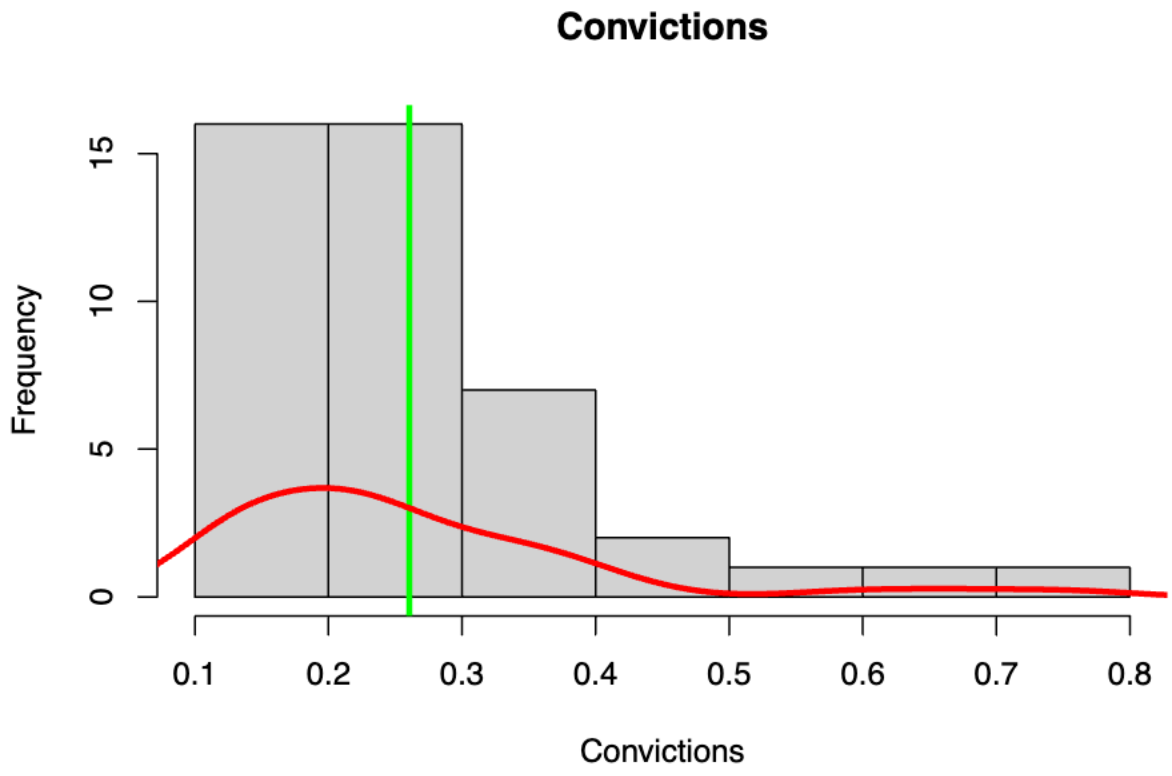
```
reg.mod <- lm(convictions ~ income, data = MurderRates)

cov1 <- hccm(reg.mod, type="hc1") #uses the car package
coeftest(reg.mod, vcov.=cov1) #produces the robust standard errors
```

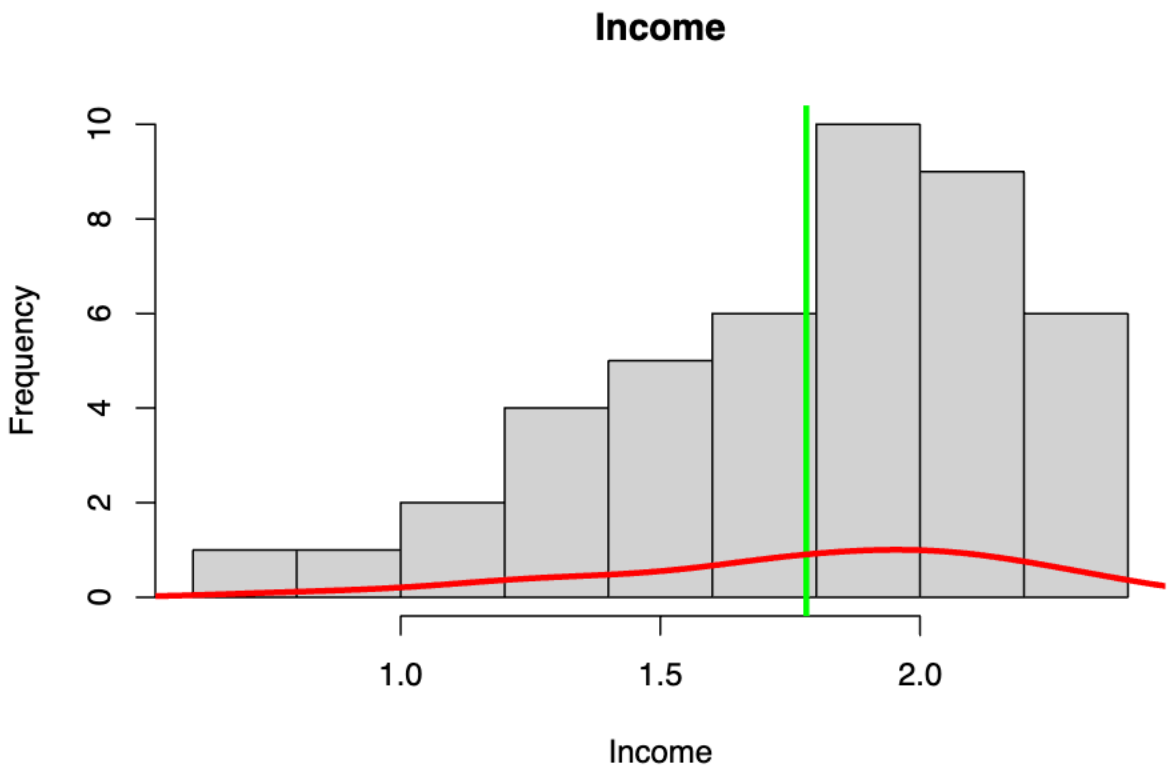
```
##
## t test of coefficients:
##
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.218569   0.060015  3.6419 0.0007371 ***
## income      0.023532   0.031770  0.7407 0.4629982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#c
hist(MurderRates$convictions, xlab = 'Convictions', ylab = 'Frequency', main = 'Convictions')
abline(v = mean(MurderRates$convictions), col='green', lwd = 3)
lines(density(MurderRates$convictions), col = 'red', lwd = 3)
```




```r
hist(MurderRates$income, xlab = 'Income', ylab = 'Frequency', main = 'Income')
abline(v = mean(MurderRates$income), col='green', lwd = 3)
lines(density(MurderRates$income), col = 'red', lwd = 3)
```




```r
#d
model <- lm(convictions ~ income, data = MurderRates)
summary(model)
```

```
##
## Call:
## lm(formula = convictions ~ income, data = MurderRates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15316 -0.09268 -0.04001  0.05578  0.49513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.21857    0.10044   2.176   0.0352 *
## income       0.02353    0.05508   0.427   0.6714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1431 on 42 degrees of freedom
## Multiple R-squared:  0.004326,   Adjusted R-squared:  -0.01938
## F-statistic: 0.1825 on 1 and 42 DF,  p-value: 0.6714
```

```r
summary(model)$coefficient
```

```
##               Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 0.21856903 0.10044381 2.176033 0.03522805
## income      0.02353194 0.05508465 0.427196 0.67141878
```

3. Estimate a multiple linear regression model that includes main effects only (i.e. no interactions or higher order terms). This is our baseline model.

(a) Comment on the statistical and economic significance of your individual estimates and provide an interpretation of the estimates obtained. Include any anomalies present if any such as unrealistic magnitudes, unexpected signs, etc.

There is no evidence that the variables explain the variation in convictions. The p-value of $0.2372 > 0.05$ so we fail to reject that the variables are different from 0.

(b) Comment on the overall fit of the model and how 1(d) might interfere with this. Comment also on the overall statistical significance of the model

The variables are not statistically significant.

```
model <- lm(convictions ~ income + southern + noncauc + lfp , data = MurderRates)
summary(model)
```

```
##
## Call:
## lm(formula = convictions ~ income + southern + noncauc + lfp,
##     data = MurderRates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19279 -0.10372 -0.00639  0.06312  0.41478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.94460    0.53053   1.780   0.0828 .
## income       0.04658    0.09698   0.480   0.6337
## southernyes  0.02016    0.07168   0.281   0.7800
## noncauc     -0.22807    0.30665  -0.744   0.4615
## lfp         -0.01413    0.01141  -1.239   0.2229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1426 on 39 degrees of freedom
## Multiple R-squared:  0.08094,    Adjusted R-squared:  -0.01332
## F-statistic: 0.8587 on 4 and 39 DF,  p-value: 0.4972
```

4. Test the model in (3) for multicollinearity using VIF. Based on this test remove the appropriate variables and estimate a new regression model based on these findings. Be sure to justify your reason/criteria for removal.

The VIF results gives us no cause to believe that there is no multicollinearity because the results are less than 10.

```r
library(car)

model <- lm(convictions ~ income + southern + noncauc + lfp , data = MurderRates)
summary(model)
```

```
##
## Call:
## lm(formula = convictions ~ income + southern + noncauc + lfp,
##     data = MurderRates)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19279 -0.10372 -0.00639  0.06312  0.41478
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.94460    0.53053   1.780   0.0828 .
## income       0.04658    0.09698   0.480   0.6337
## southernyes  0.02016    0.07168   0.281   0.7800
## noncauc     -0.22807    0.30665  -0.744   0.4615
## lfp         -0.01413    0.01141  -1.239   0.2229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1426 on 39 degrees of freedom
## Multiple R-squared:  0.08094,    Adjusted R-squared:  -0.01332
## F-statistic: 0.8587 on 4 and 39 DF,  p-value: 0.4972
```

```r
# Calculate VIF
vif_results <- car::vif(model)

# Print the VIF for each predictor
print(vif_results)
```

```
##   income southern  noncauc      lfp
## 3.118394 2.496359 2.582296 1.692073
```

5. Using AIC or Schwartz Criterion, determine which subset of predictors you will keep and generate a new model. Comment on the performance of this model compared to the one in (3)

The lower value of AIC the better the fit of the model, thus since the AIC is -39.8189 we an conclude that we can keep all predictors. All 4 variables create the best fit model.

```r
model <- lm(convictions ~ income + southern + noncauc + lfp , data = MurderRates)

# Assuming you have a multiple regression model called 'model'

# Load necessary library
library(stats)

# Calculate AIC for the model
aic <- AIC(model)
```

```r
# Print the AIC value
print("AIC:")
```

```
## [1] "AIC:"
```

```r
print(aic)
```

```
## [1] -39.81289
```

```r
model <- lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates)

# Load necessary libraries
library(stats)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Predictors in the model
predictors <- c("income", "southern", "noncauc", "lfp")

# Generate all possible subsets of predictors
all_subsets <- lapply(1:length(predictors), function(k) combn(predictors, k))

# Calculate AIC for each subset
aic_values <- sapply(all_subsets, function(subset) {
  if (length(subset) > 0) {
    # Fit a model using the current subset
    formula <- paste("convictions ~", paste(subset, collapse = " + "))
    model <- lm(formula, data = MurderRates)

    # Calculate AIC for the model
    AIC(model)
  } else {
    Inf  # No predictors in the subset, set AIC to infinity
  }
})
```

```r
# Find the subset with the lowest AIC
best_subset <- all_subsets[[which.min(aic_values)]]

# Print the best subset and its AIC
print("Best Subset:")
```

```
## [1] "Best Subset:"
```

```r
print(best_subset)
```

```
##      [,1]     [,2]       [,3]      [,4]
## [1,] "income" "southern" "noncauc" "lfp"
```

```r
print("AIC for Best Subset:")
```

```
## [1] "AIC for Best Subset:"
```

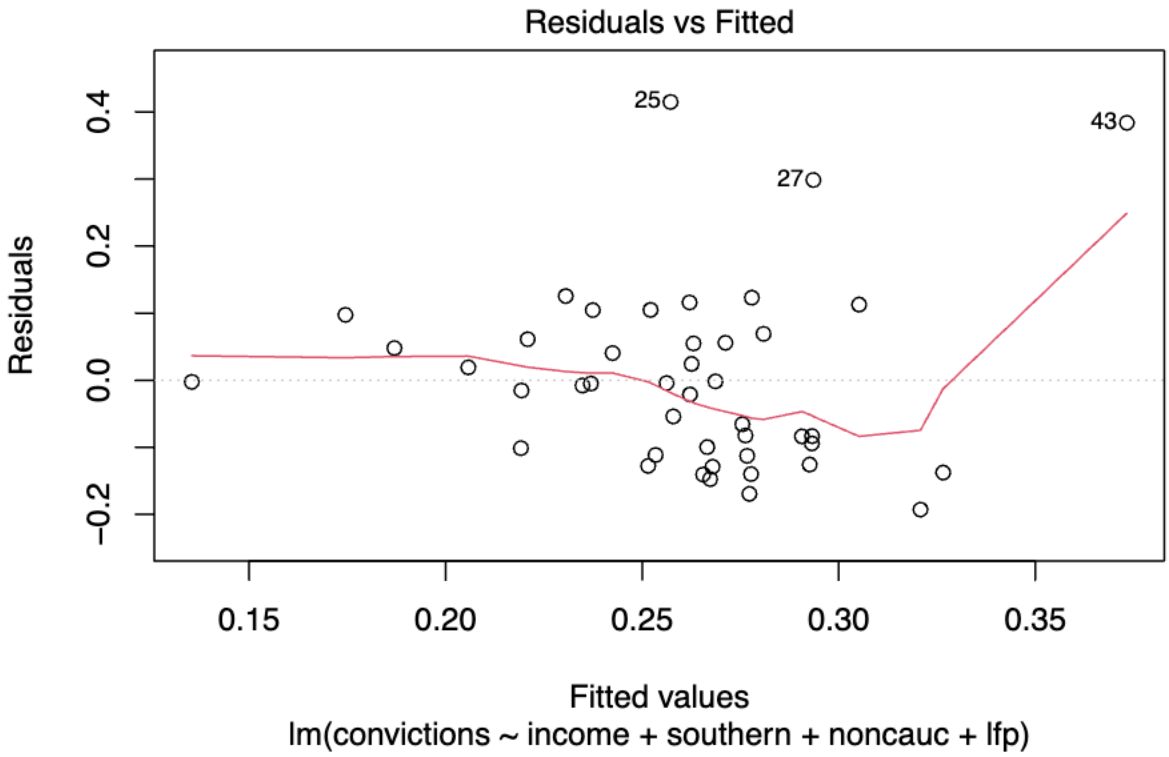
```r
print(min(aic_values))
```

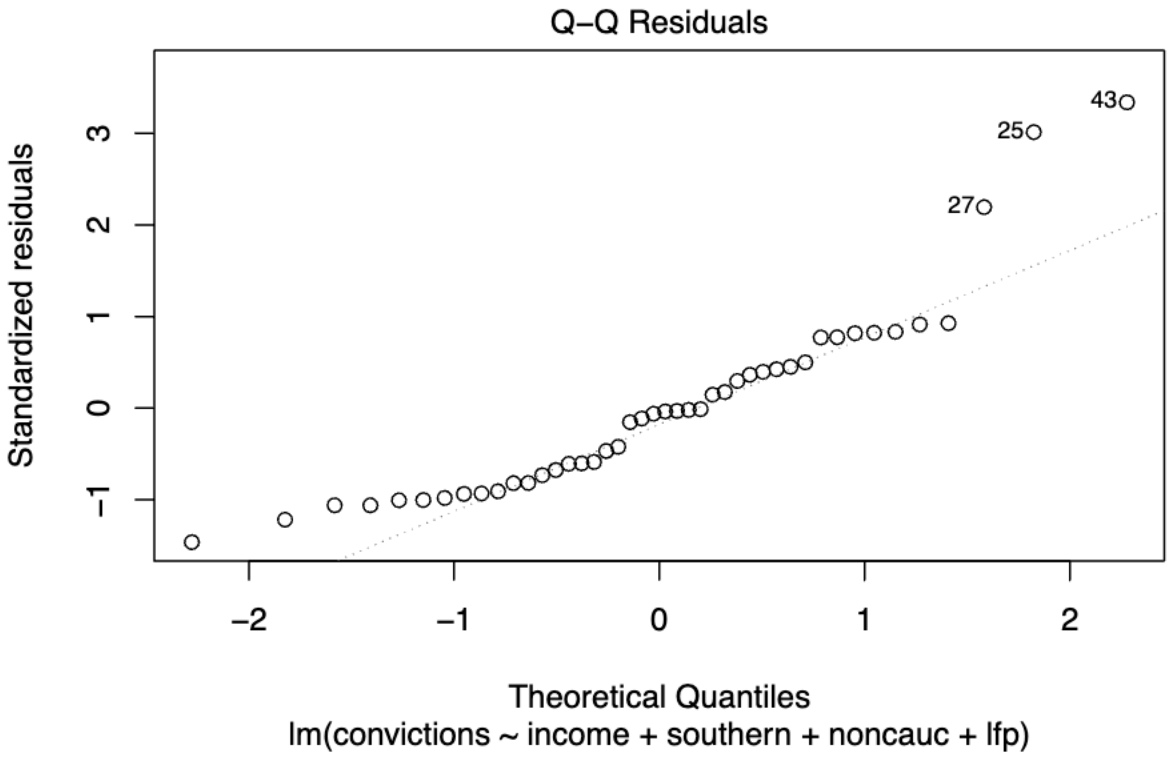
```
## [1] -39.81289
```

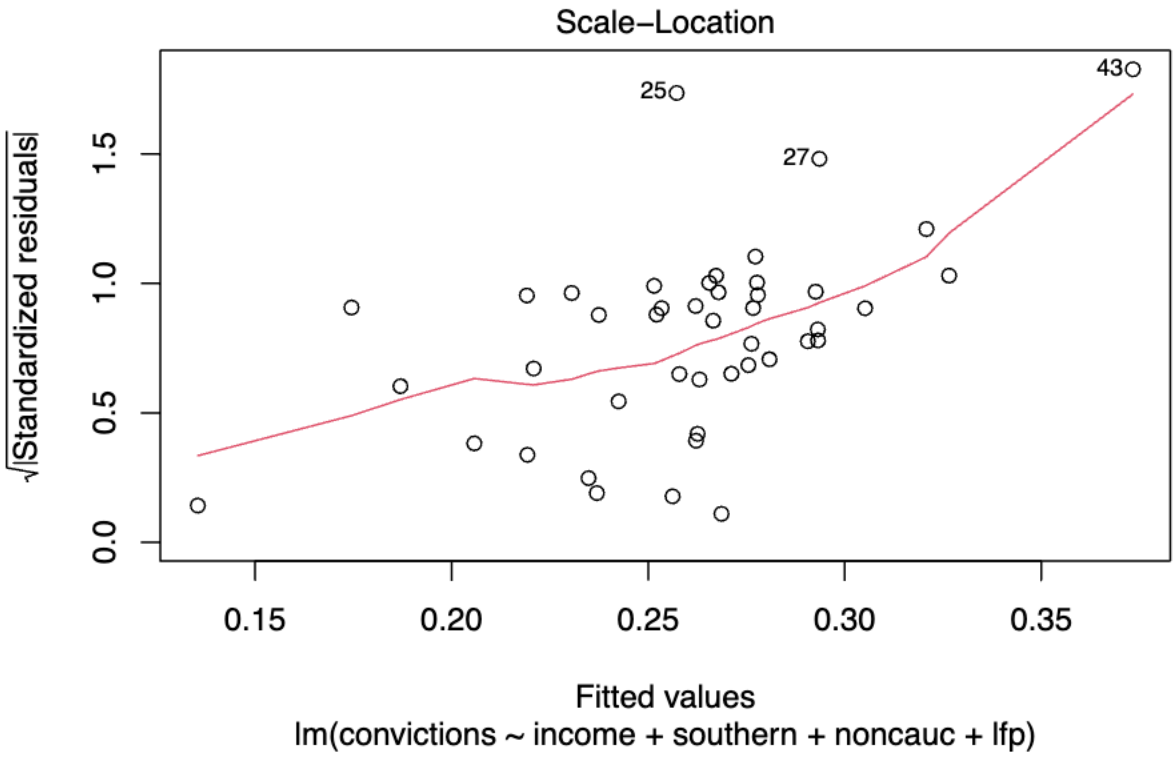
6. Using the model in (5) plot the residuals versus its fitted values, ^y and comment on your results.

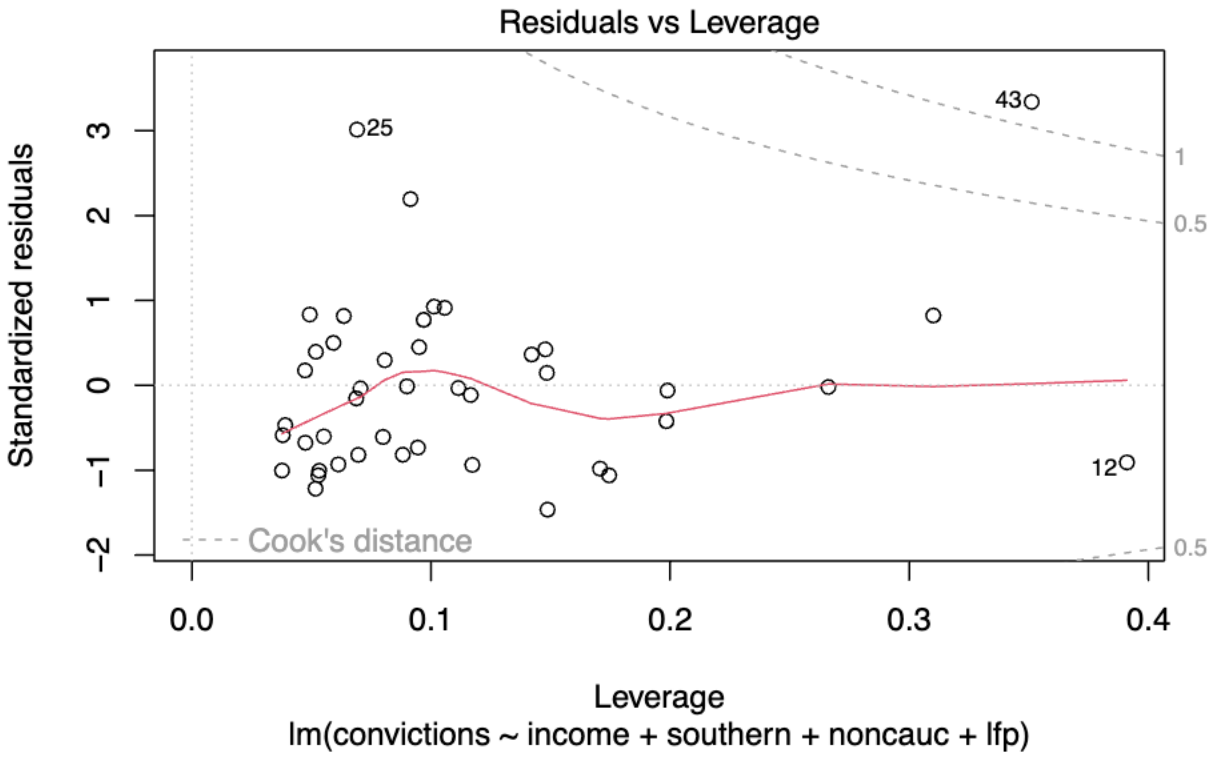
The red line starts off linear towards the left but the linearity does not hold past 0.25. The spread of the residuals increases at 0.30 indicating heteroskedasticity.

```r
library(mlbench)
data("MurderRates")
model <- lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates)
plot(lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates))
```


Residuals vs Fitted
25
43
27
Residuals
0.4
0.2
0.0
−0.2
0.15
0.20
0.25
0.30
0.35
Fitted values
lm(convictions ~ income + southern + noncauc + lfp)


Q–Q Residuals
Standardized residuals
Theoretical Quantiles
lm(convictions ~ income + southern + noncauc + lfp)
43
25
27
−2
−1
0
1
2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.15
0.20
0.25
0.30
0.35
25
27
43
Fitted values
lm(convictions ~ income + southern + noncauc + lfp)



7. Perform a RESET test on the model in (5) and comment on the results

Since the p-value, $0.003395 < 0.05$, we can reject the null that we used the correct functional form. A significant result indicates that this model may be misspecified. Thus, our functional form is incorrect and our model might suffer from omitted variables.

```
model <- lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates)
resettest(model, power = 2:3, type = c("fitted", "regressor",
  "princomp"), data = list(MurderRates), vcov = NULL)
```

```
##
##  RESET test
##
## data:  model
## RESET = 6.6558, df1 = 2, df2 = 37, p-value = 0.003395
```

8. Using the appropriate method learnt in class, test the model in (4) for heteroskedastic- ity and comment on the conclusion. If it is present, correct the model before moving on. Based on the results in (c) or (d), this might be helpful in transforming the model in the event that its functional form presents an issue.

Since the p-value is $0.106 > 0.05$, we fail to reject the null that there is heteroskedasticity. We can assume our coefficients are equal to 0. Thus, we do not have to correct the model.

```r
model <- lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates)

#perform White's test
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 7.6323, df = 4, p-value = 0.106
```

9.Using a combination of the results from the previous steps, estimate a model based on your findings which includes interaction terms or higher power terms (if necessary). You may need to use forward or backward selection for this. Comment on the per- formance of this model. compared to your other models. Make sure to use AIC and Schwartz criterion for model comparison.

When estimating a new model using the cross-product of lfp and income, we got a lower pvalue result indicating that this new regression has a better functional form. When comparing AICs between models, results show that the new AIC is lower than the previous AIC. The original model's Schwartz criterion has a lower value of -32.89194 compared to the new model which has a value of -29.37393 giving us opposing results. For the BIC test, the original equation is better but for the AIC test, the new function is better.

```r
# For the White's res eqn we'll use x and x^2
model <- lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates)
alpha <- 0.05
ressq <- resid(model)^2
# The test equation:
modres <- lm(convictions~income+ southern + noncauc + I(lfp*income))
summary(modres)
```

```
##
## Call:
## lm(formula = convictions ~ income + southern + noncauc + I(lfp *
##     income))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18784 -0.10433 -0.01057  0.06952  0.41177
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.192829   0.193682   0.996    0.326
## income            0.492789   0.393115   1.254    0.217
## southernyes       0.022205   0.071170   0.312    0.757
## noncauc          -0.221465   0.304799  -0.727    0.472
## I(lfp * income) -0.008356   0.006284  -1.330    0.191
##
## Residual standard error: 0.1422 on 39 degrees of freedom
## Multiple R-squared:  0.08622,    Adjusted R-squared:  -0.0075
## F-statistic:  0.92 on 4 and 39 DF,  p-value: 0.4621
```

```r
# Calculate AIC for the model
aic <- AIC(modres)

# Print the AIC value
print("AIC:")
```

```
## [1] "AIC:"
```

```r
print(aic)
```

```
## [1] -40.06642
```

```r
# Fit the model
model <- lm(convictions ~ income + southern + noncauc + lfp, data = MurderRates)

# Calculate log-likelihood
log_likelihood <- logLik(model)

# Get the number of parameters (including intercept)
num_parameters <- length(coef(model))

# Get the sample size
n <- nrow(model$model)

# Calculate BIC
BIC <- -2 * log_likelihood + num_parameters * log(n)

# Print BIC
print(BIC)
```

```
## 'log Lik.' -32.89194 (df=6)
```

```r
# Fit the modres
model <- lm(convictions ~ income + southern + noncauc + lfp + I(lfp*income), data = MurderRates)

# Calculate log-likelihood
log_likelihood <- logLik(model)

# Get the number of parameters (including intercept)
num_parameters <- length(coef(model))

# Get the sample size
n <- nrow(model$model)

# Calculate BIC
BIC <- -2 * log_likelihood + num_parameters * log(n)

# Print BIC
print(BIC)
```

```
## 'log Lik.' -29.37393 (df=7)
```

10. Provide a short 1 paragraph summary of your overall conclusion, findings, and recom- mendations not previously stated above

Through these results, we found that the model that included all 4 variables including income, noncauc, lfp, and southern was a good fit. However, we found that including the cross-product (lfp*income) further improved the model. Our results showed that heteroskedasticity was not present in the model. However, we found opposing results between the AIC and BIC tests when comparing the original model versus the new model.