Homework 5

Jun Ryu, UID: 605574052

2023-05-05

```r
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(dplyr)
```

Components of ANOVA Table

a)

RSS is $\hat{\sigma}^2 * (n-2)$, where $n-2$ represents the degrees of freedom. We can find $\hat{\sigma}$ from the table where it says residual standard error.

```r
df <- 33
rss <- 2.418^2 * df
rss
```

```
## [1] 192.9419
```

b)

We have $F = \frac{SS_{reg}/1}{RSS/df}$, so $SS_{reg} = \frac{F*RSS}{df}$.

```r
ss_reg <- 87.17*rss/df
ss_reg
```

```
## [1] 509.6589
```

c)

Mean SSreg is just $\frac{SS_{reg}}{df}$.

```
mean_ss <- ss_reg/df
mean_ss
```

```
## [1] 15.44421
```

d)

Total SS is $SS_{reg} + RSS$.

```
total_ss <- ss_reg + rss
total_ss
```

```
## [1] 702.6008
```

e)

Correlation coefficient can be found by $\sqrt{\frac{SS_{reg}}{SYY}}$, where SYY is the total SS.

```
r <- sqrt(ss_reg/total_ss)
r
```

```
## [1] 0.8516977
```

Question 1

```
armspan <- read.csv("armspans2022_gender.csv")
head(armspan)
```

```
##   height armspan is.female compmother     compfather
## 1  74.00    76.0         0     Taller         Taller
## 2  65.00    65.0         0     Taller About the same
## 3  60.00    53.0         1    Shorter        Shorter
## 4  69.75    69.0         0     Taller About the same
## 5  70.00    72.0         0     Taller About the same
## 6  68.00    70.5         0     Taller        Shorter
```

a)

```
# 1 in the is.female column represents a female
mean(armspan$is.female)
```

```
## [1] 0.3478261
```

b)

```
model <- lm(armspan ~ is.female, data = armspan)
summary(model)
```

```
##
## Call:
## lm(formula = armspan ~ is.female, data = armspan)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7586 -2.0248  0.2414  2.2414  8.2414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.7586     0.7399  94.284  < 2e-16 ***
## is.female    -7.7338     1.2408  -6.233 1.68e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.984 on 43 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4746, Adjusted R-squared:  0.4624
## F-statistic: 38.85 on 1 and 43 DF,  p-value: 1.676e-07
```

The intercept is 69.7586. This represents that the average armspan length for males in the dataset (is.female = 0) is about 69.76 inches.

c)

The slope is -7.7338. This represents that on average, females in the dataset have an armspan length that is 7.7338 inches shorter than that of males.

d)

The t-testistic and the p-value for the slope is testing our null hypothesis $H_0 : \beta_1 = 0$ and our alternative hypothesis $H_a : \beta_1 \neq 0$. In this context, it tests whether there is a statistically significant difference between the average armspan length for males and females.

Question 2

```
iowa <- read.delim("iowatest.txt", header = T)
head(iowa)
```

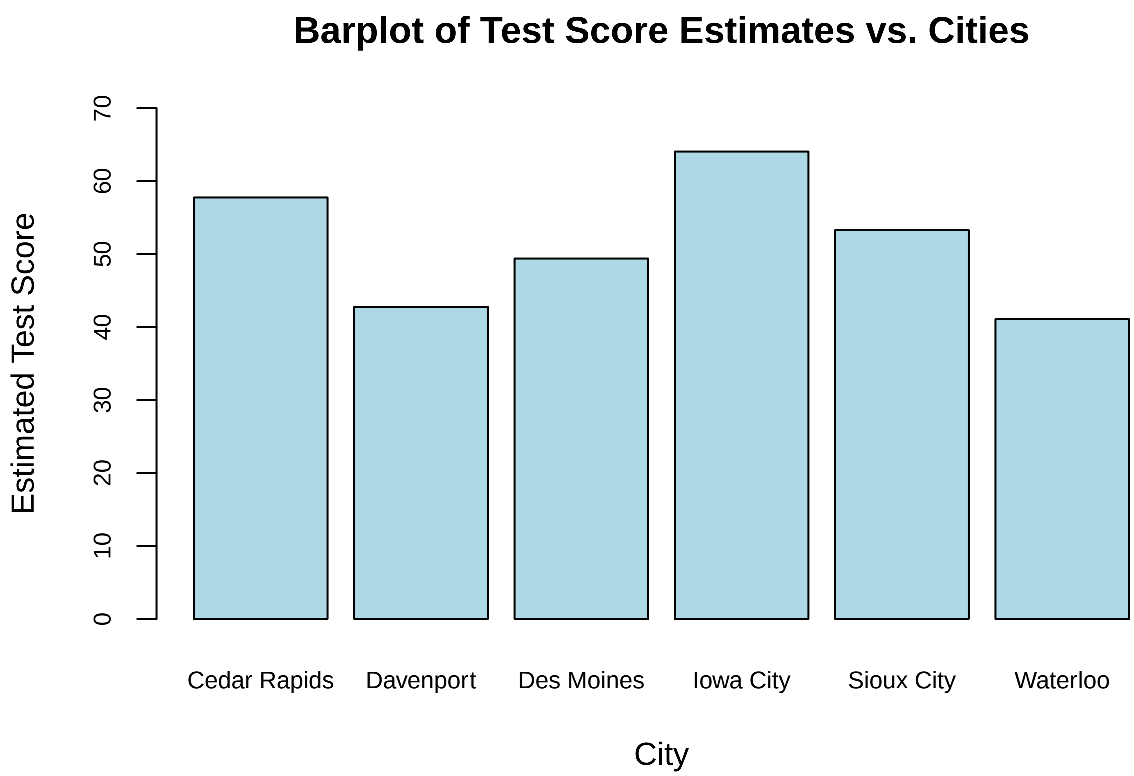
```
##       School Poverty Test      City
## 1 Coralville      20   65 Iowa City
## 2      Hills      42   35 Iowa City
## 3     Hoover      10   84 Iowa City
```

```
## 4       Horn       5   83 Iowa City
## 5   Kirkwood      34   49 Iowa City
## 6      Lemme      17   69 Iowa City
```

```r
model2 <- lm(Test ~ factor(City), data = iowa)
summary(model2)
```

```
##
## Call:
## lm(formula = Test ~ factor(City), data = iowa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.059 -10.286   0.227   9.605  34.929
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                57.762      2.992  19.308  < 2e-16 ***
## factor(City)Davenport     -14.989      4.182  -3.584 0.000481 ***
## factor(City)Des Moines     -8.367      3.728  -2.245 0.026524 *
## factor(City)Iowa City       6.297      4.473   1.408 0.161619
## factor(City)Sioux City     -4.476      4.231  -1.058 0.292060
## factor(City)Waterloo      -16.690      4.730  -3.529 0.000582 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.71 on 127 degrees of freedom
## Multiple R-squared:  0.225,  Adjusted R-squared:  0.1945
## F-statistic: 7.373 on 5 and 127 DF,  p-value: 4.238e-06
```

```r
estimates <- summary(model2)$coefficients
# since all estimates are relative to the first estimate:
estimates[2:6] <- estimates[2:6] + estimates[1]
cities <- sort(unique(iowa$City))
par(cex.axis = 0.75)
barplot(estimates[,1], names.arg = cities, xlab = "City", ylab = "Estimated Test Score",
        col = "lightblue", main = "Barplot of Test Score Estimates vs. Cities",
        ylim = c(0,70))
```



Comparing Iowa City against the other 5 cities, we notice that Iowa City does have the highest estimated test scores so we can conclude that Iowa City does outperform.

Question 3

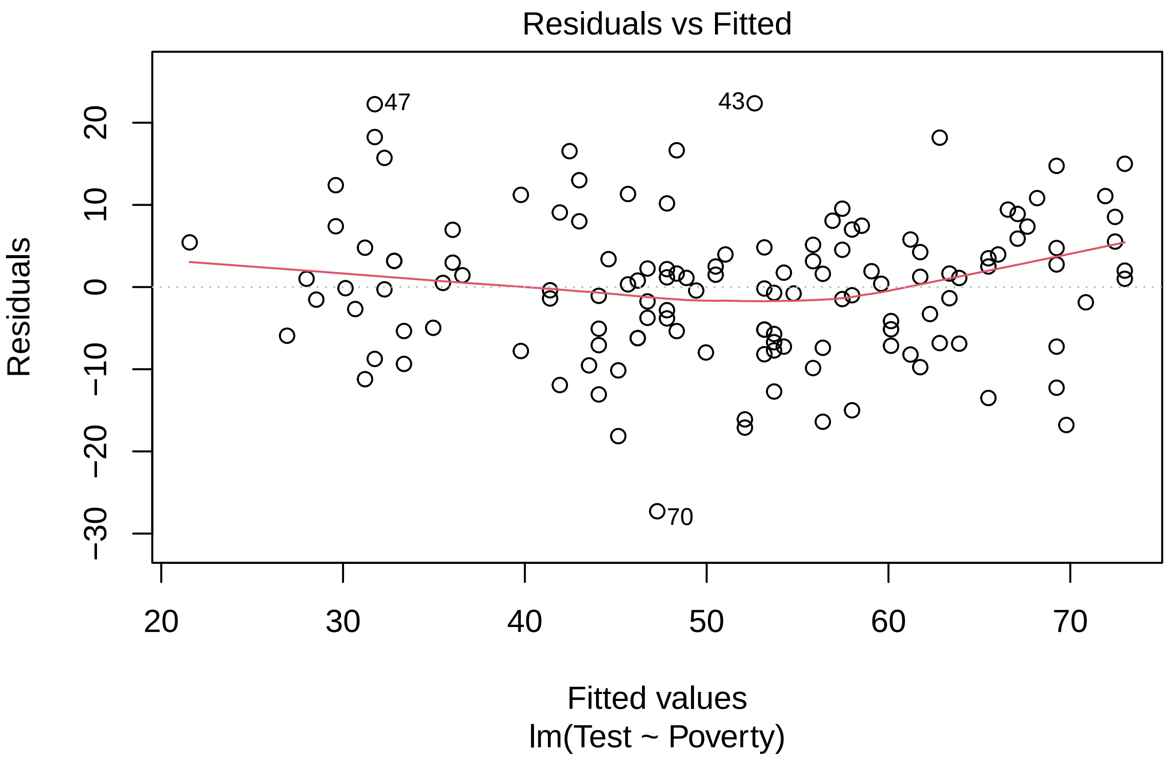
```r
model3 <- lm(Test ~ Poverty, data = iowa)
summary(model3)
```

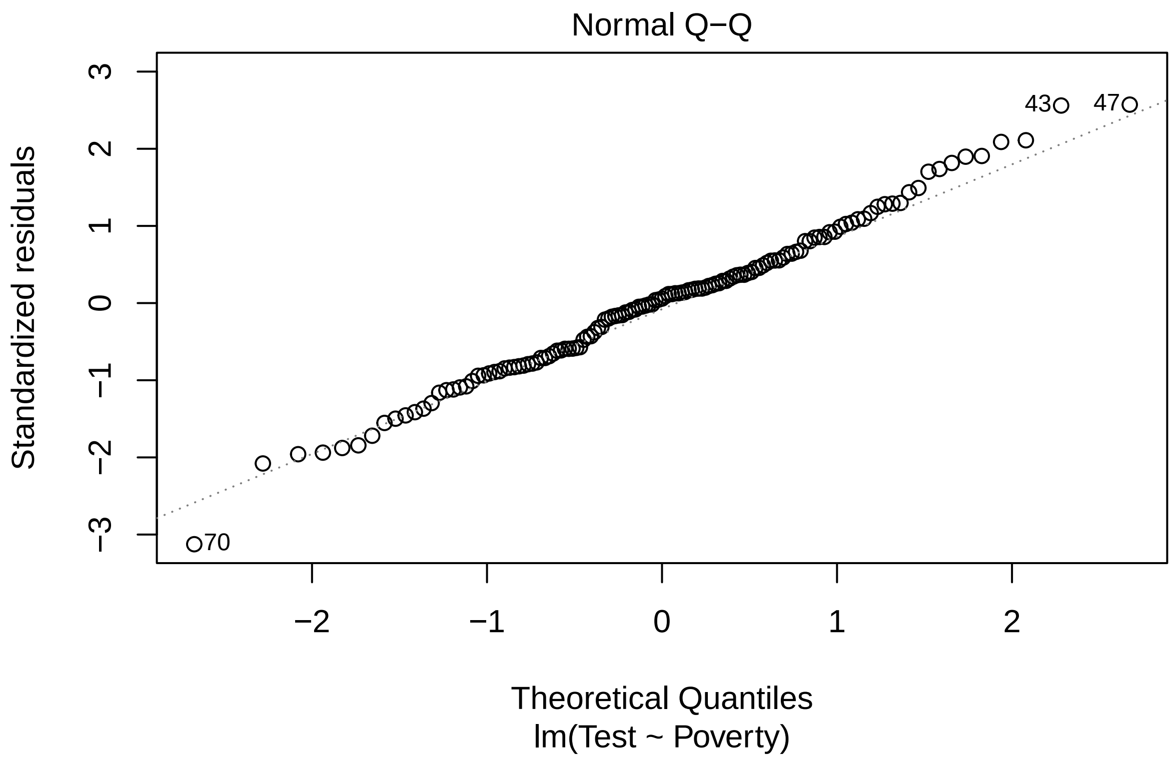
```
##
## Call:
## lm(formula = Test ~ Poverty, data = iowa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2812  -6.2097   0.5058   4.8252  22.3610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.60578    1.61325   46.25   <2e-16 ***
## Poverty     -0.53578    0.03262  -16.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.766 on 131 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6707
## F-statistic: 269.8 on 1 and 131 DF,  p-value: < 2.2e-16
```

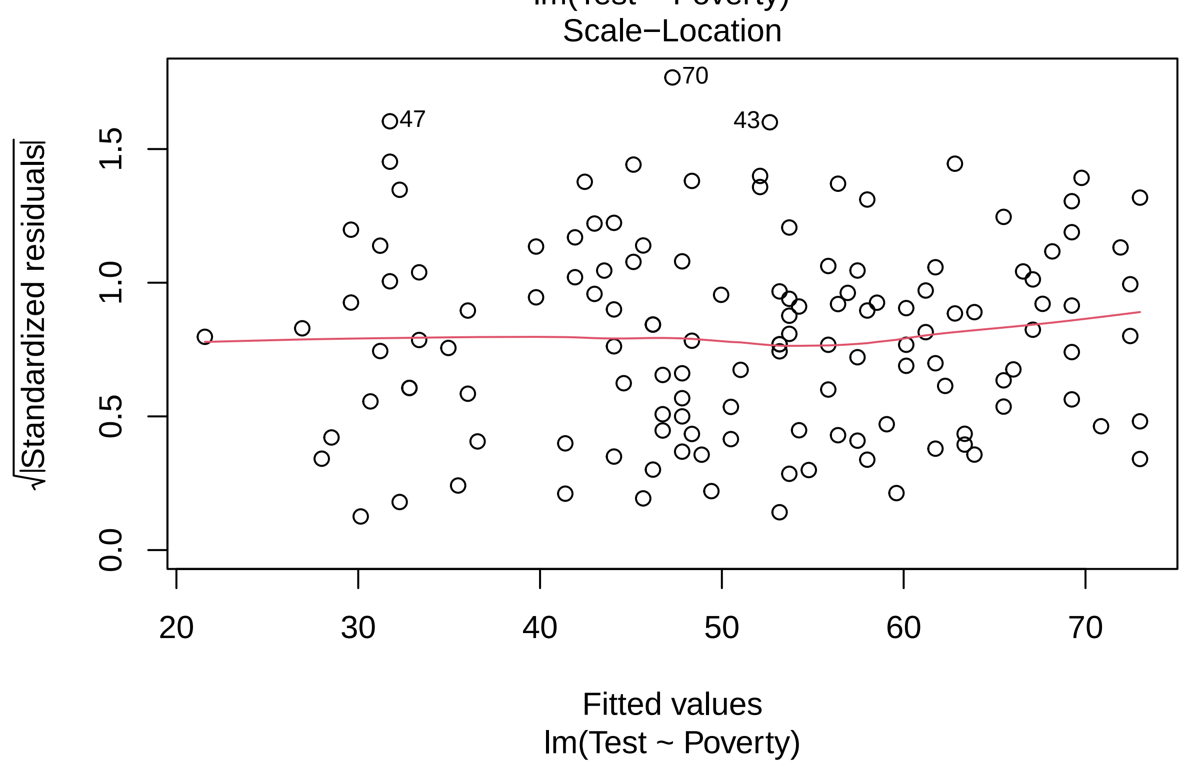
We test this by testing the hypothesis that our slope, or $\beta_1 = 0$. So, we have our null is $H_0 : \beta_1 = 0$ and our alternative is $H_a : \beta_1 \neq 0$. Looking at the summary table above, we observe that the p-value for this test is $< 2 * 10^{-16}$, which leads us to reject the null and conclude that there is evidence that poverty is associated with the test score.

Question 4

```
plot(model3, which = 1:3)
```







1) The residuals vs fitted plot shows a mostly flat red line, indicating a good linear association. Also, the red line does not create a huge “fan” shape, indicating a constant variance.

2) The qq-plot is mostly straight. Even at the ends of the plot, the points do not diverge too much from the dotted line. This indicates normality of our model.

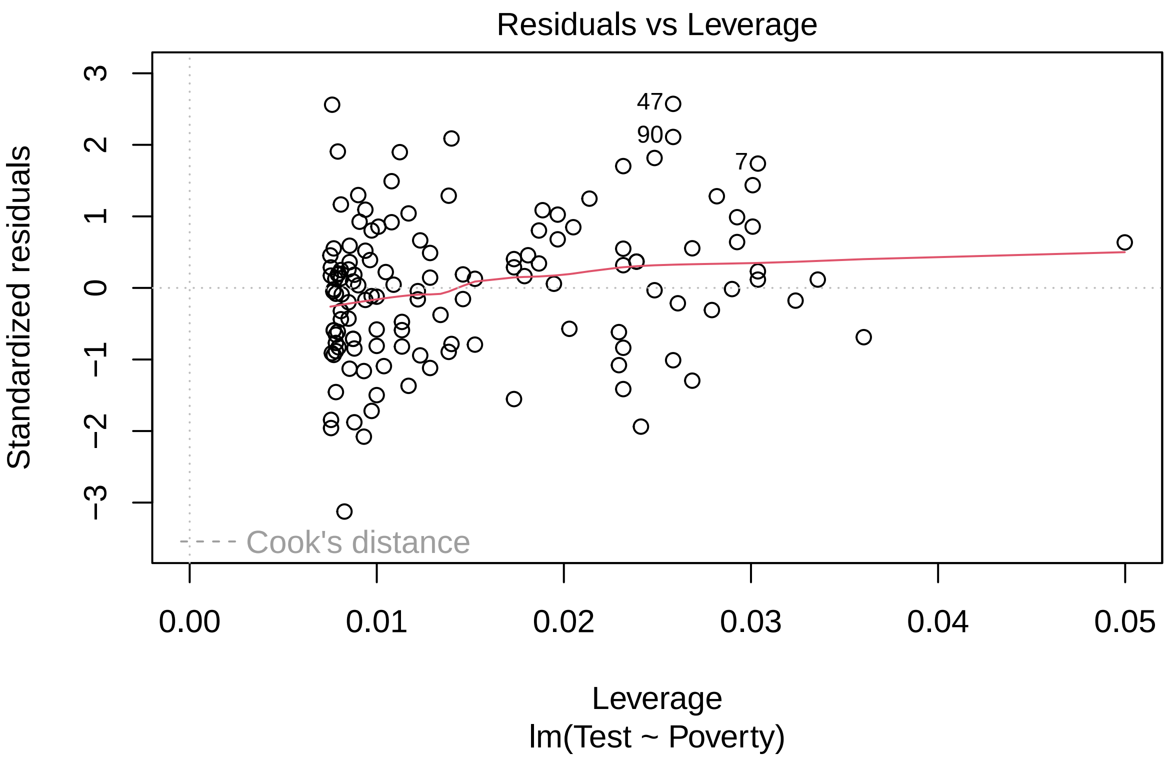
3) The scale-location plot also shows a relatively flat red line and we do not observe any trend among the

data points, indicating constant variance.

Combining results from the three plots, we see that the model is a valid one as it follows normality, homoscedasticity, and linearity.

Question 5

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



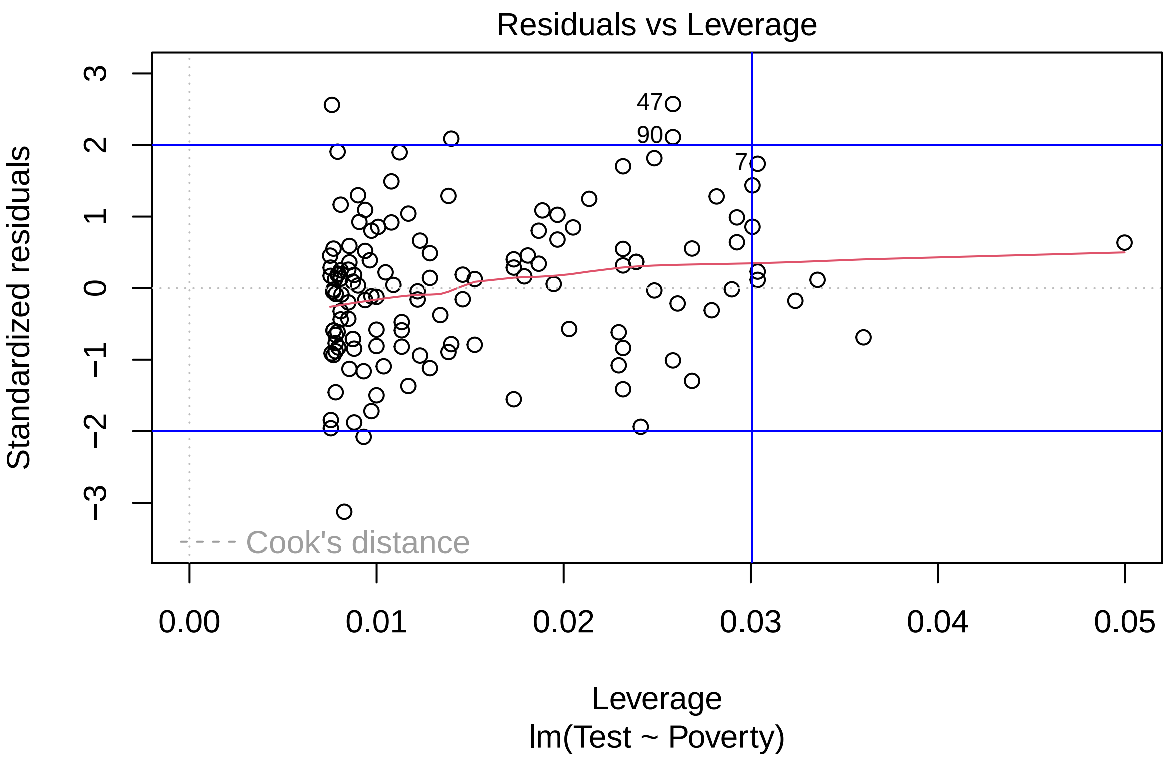

```
#hatvalues() returns a list of all observations' leverages
which.max(hatvalues(model3))
```

```
## 27
## 27
```

The point with the highest leverage is the point on the far right side. The 27th row corresponds to this leverage.

Now, a point is generally a bad leverage point if a) leverage is more than $4/n$ and b) the standard residual is outside of $(-2, 2)$.

```
plot(model3, which = 5)
n <- nrow(iowa)
abline(v = 4/n, col = "blue")
abline(h = c(-2,2), col = "blue")
```



So, we look for points that have a higher leverage than 0.03 and fall outside of $(-2, 2)$ in terms of y-axis. Looking at the residual-leverage graph and the plotted lines above, there is no such point, so there seems to be no bad leverage points for the data.

Question 6

```
summary(model3)
```

```
##
## Call:
## lm(formula = Test ~ Poverty, data = iowa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2812  -6.2097   0.5058   4.8252  22.3610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.60578    1.61325   46.25   <2e-16 ***
## Poverty     -0.53578    0.03262  -16.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.766 on 131 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6707
## F-statistic: 269.8 on 1 and 131 DF,  p-value: < 2.2e-16
```

The F-test is used to test our null $H_0 : \beta_1 = 0$ and our alternative $H_a : \beta_1 \neq 0$. Since our F-value is large and it yields a p-value of $< 2 * 10^{-16}$, using a significance level of 5%, we reject the null and reach the same conclusion as we did in Question 3 (there is evidence that poverty is associated with the test score).