Stats_101A_hw6_anna_piskun

Anna Piskun

2/4/2020

```
library(dplyr)
```

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

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

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

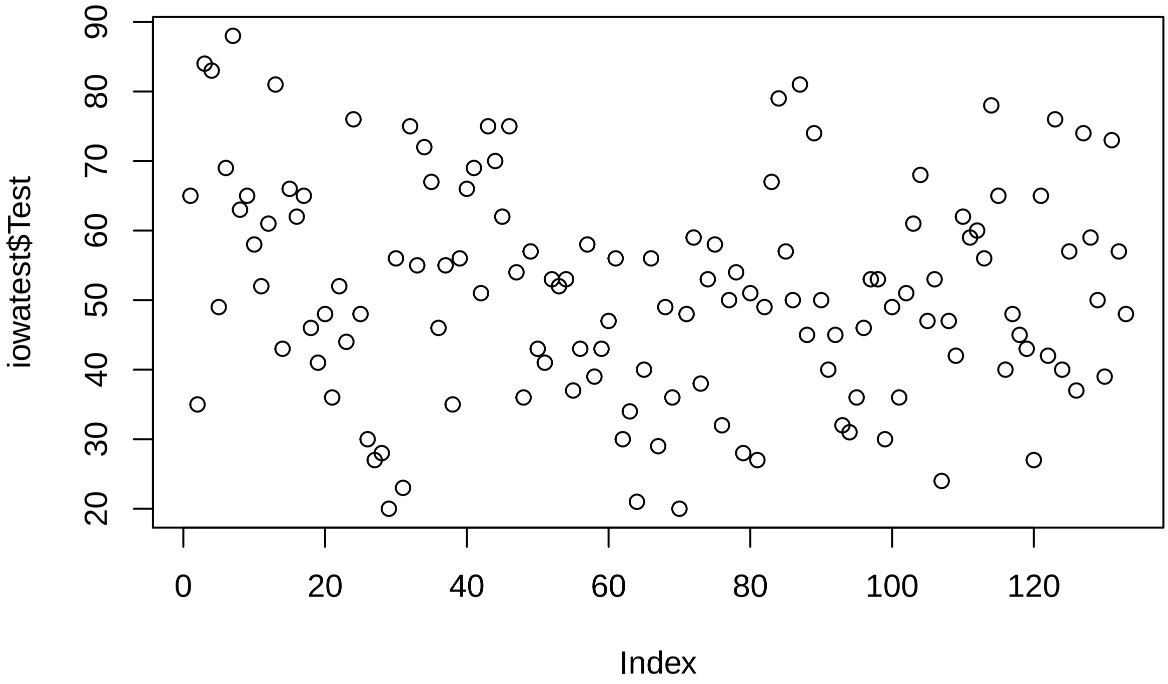
```
library(ggplot2)
```

Directions:

Question 1

Download the iowatest data. Iowa City is the home of the university of Iowa. Do schools in Iowa City out perform the rest? Answer, and provide supporting statistics and graphics (a graphic is required.)

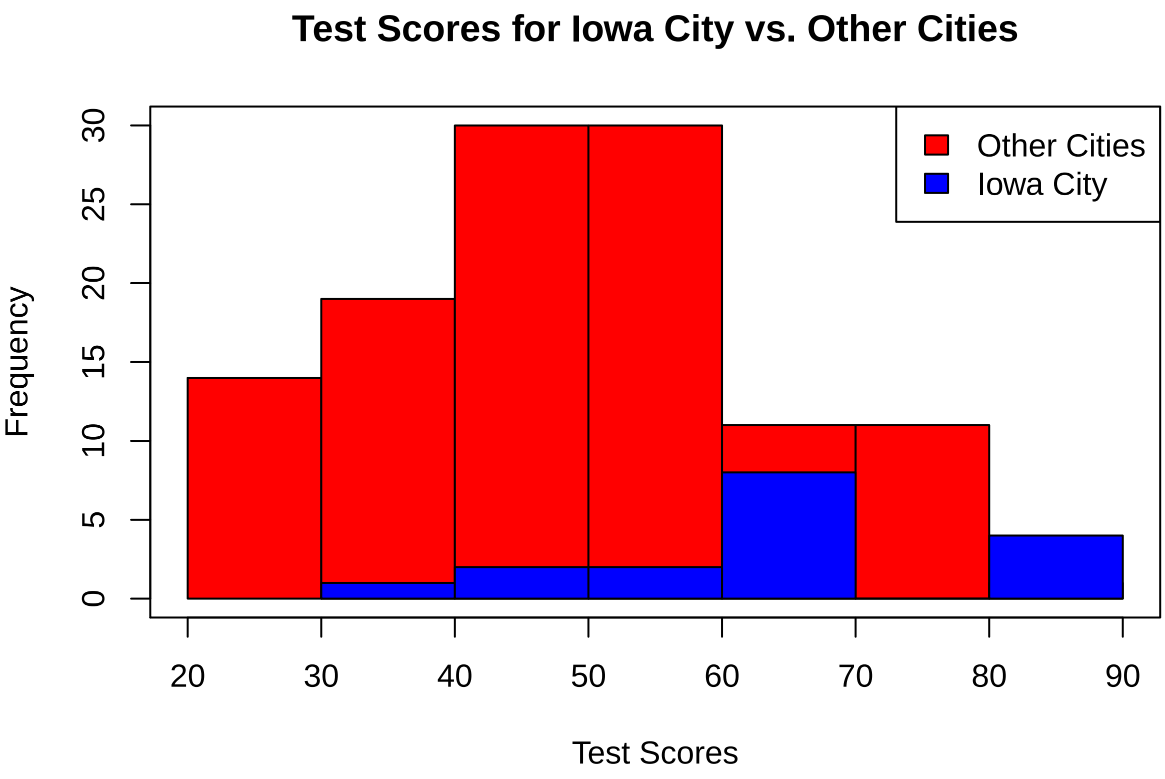
```
setwd("~/Desktop")
iowatest <- read.table("iowatest.txt", header = T, sep = "\t", fill = FALSE)

plot(iowatest$Test)
```



```
iowa_city <- dplyr::filter(iowatest, City == "Iowa City")
other_cities <- dplyr::filter(iowatest, City != "Iowa City")

t.test(x = iowa_city$Test, y = other_cities$Test, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  iowa_city$Test and other_cities$Test
## t = 3.9071, df = 20.99, p-value = 0.0004058
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  8.228843      Inf
## sample estimates:
## mean of x mean of y
##  64.05882  49.35345
```

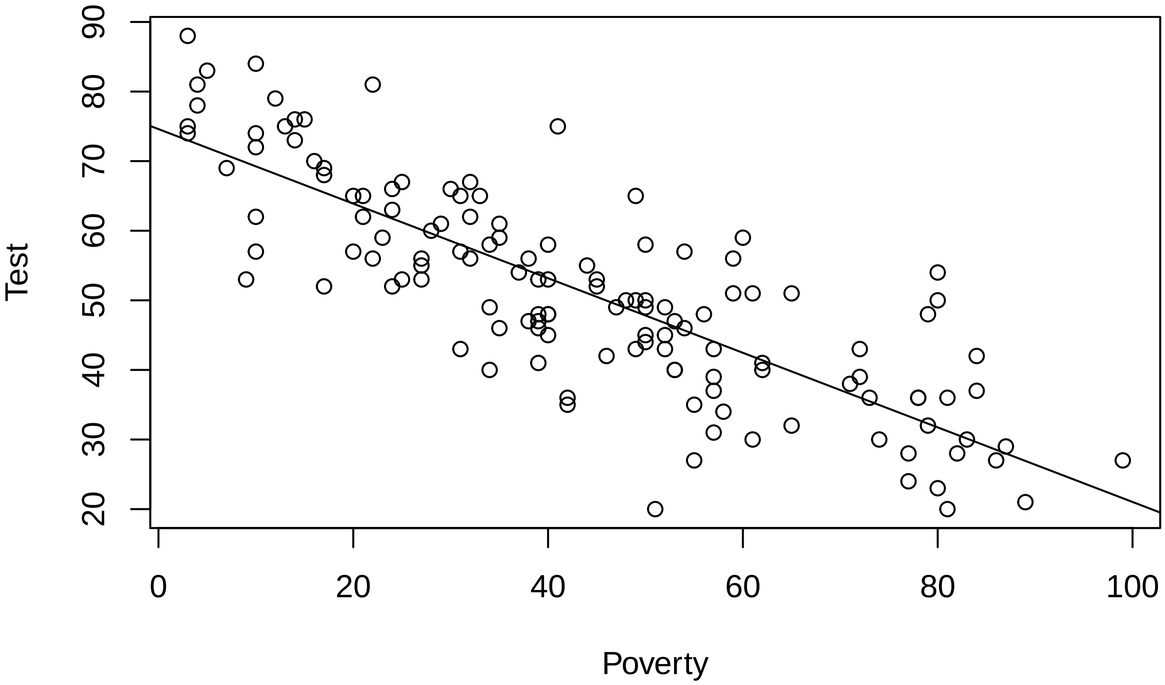
```
#Looking at our hypothesis test, since the p-value is less than 0.05 we reject the null hypothesis ther

#graphic
hist(other_cities$Test, col="red", xlab= "Test Scores", ylab= "Frequency", main= "Test Scores for Iowa
hist(iowa_city$Test, col="blue", add=TRUE)
legend("topright", c("Other Cities", "Iowa City"), fill=c("red", "blue"))
box()
```



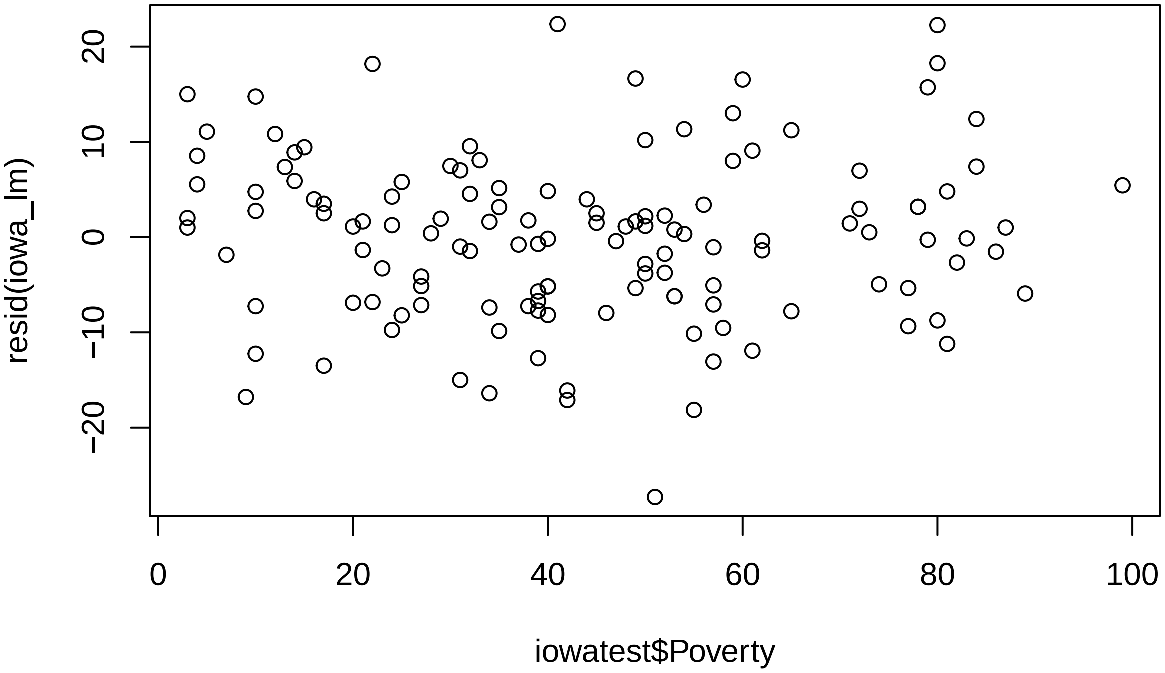
Question 2

Test scores are meant to reflect the success of a school's academic program. But many critics point out that factors other than academic success can influence a score. In particular, a school's score might be merely a reflection of the wealth of the student body. Address this issue by fitting a regression line to predict school test score from poverty score. Is there evidence that poverty is associated with the test score?

```r
plot(Test ~ Poverty, data = iowatest)
iowa_lm <- lm(Test~Poverty, data = iowatest)
abline(iowa_lm)
```



```
plot(iowatest$Poverty, resid(iowa_lm))
```



```
cor(iowatest$Test, iowatest$Poverty)
```

```
## [1] -0.8204564
```

```
summary(iowa_lm)
```

```
##
## Call:
## lm(formula = Test ~ Poverty, data = iowatest)
##
```

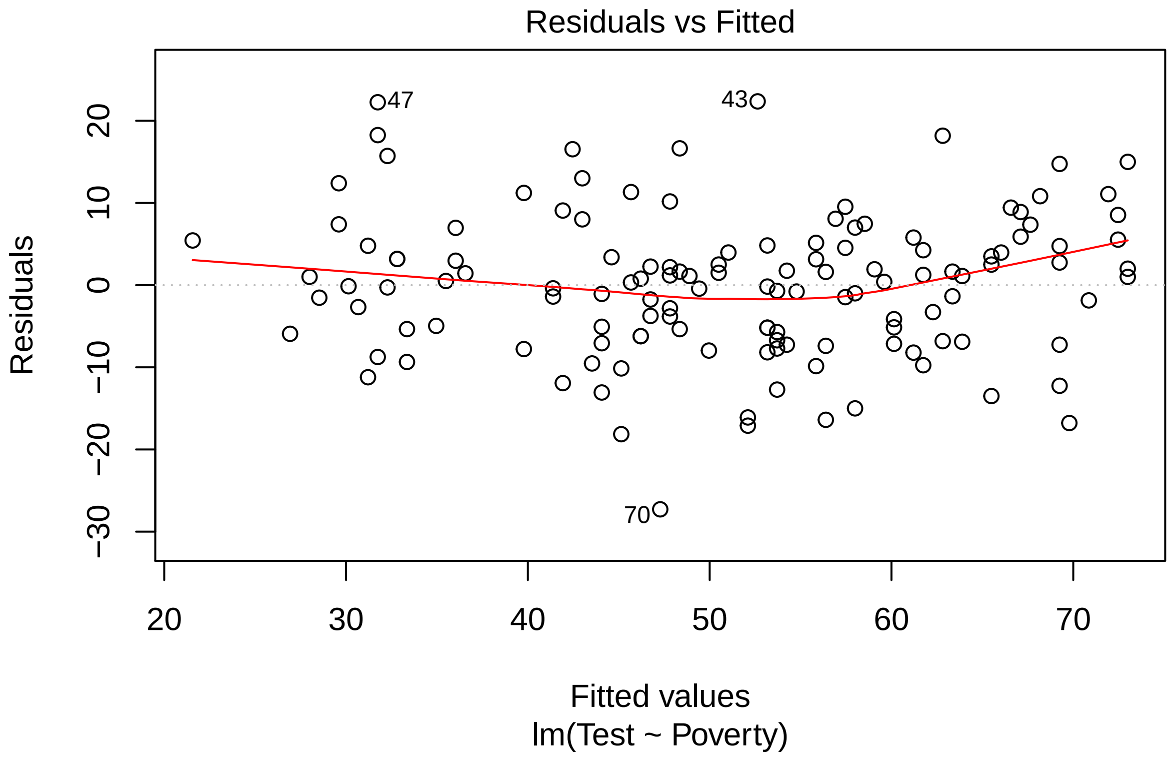
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2812  -6.2097   0.5058   4.8252  22.3610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.60578    1.61325   46.25   <2e-16 ***
## Poverty     -0.53578    0.03262  -16.43   <2e-16 ***
## ---
## 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
```

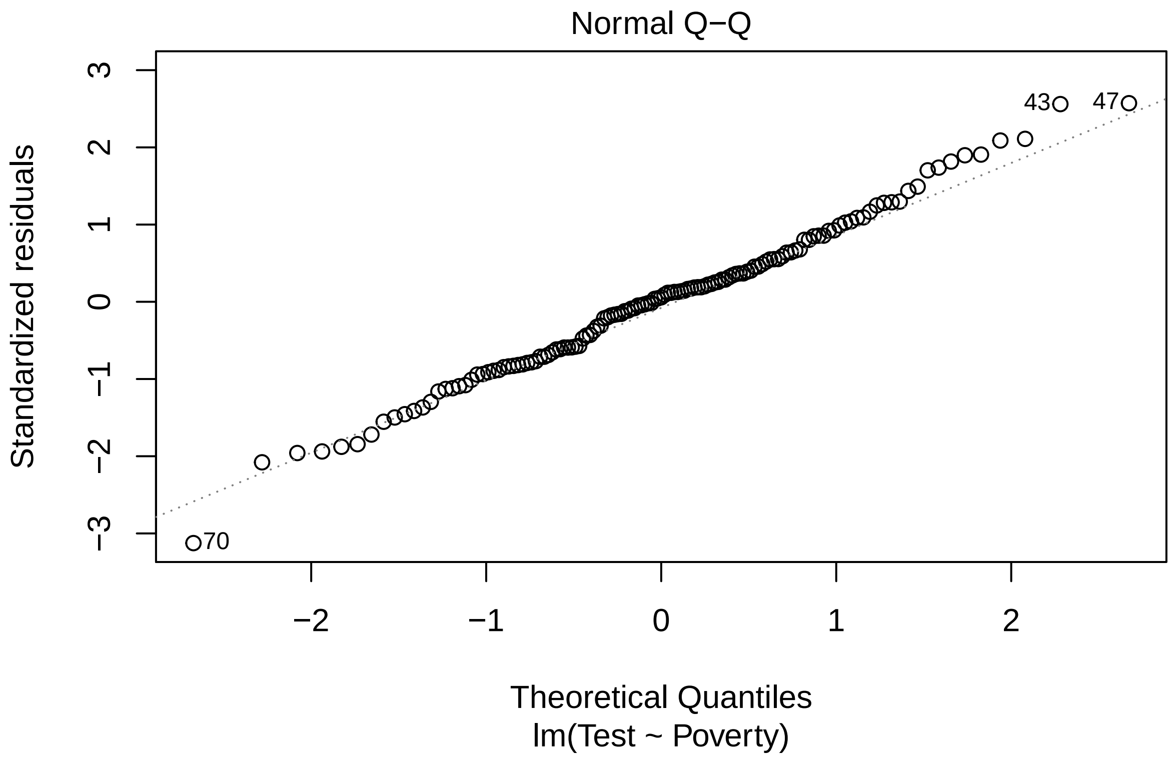
```
#Yes, there is evidence that poverty is associated with test scores. There is a strong, negative, linea
```

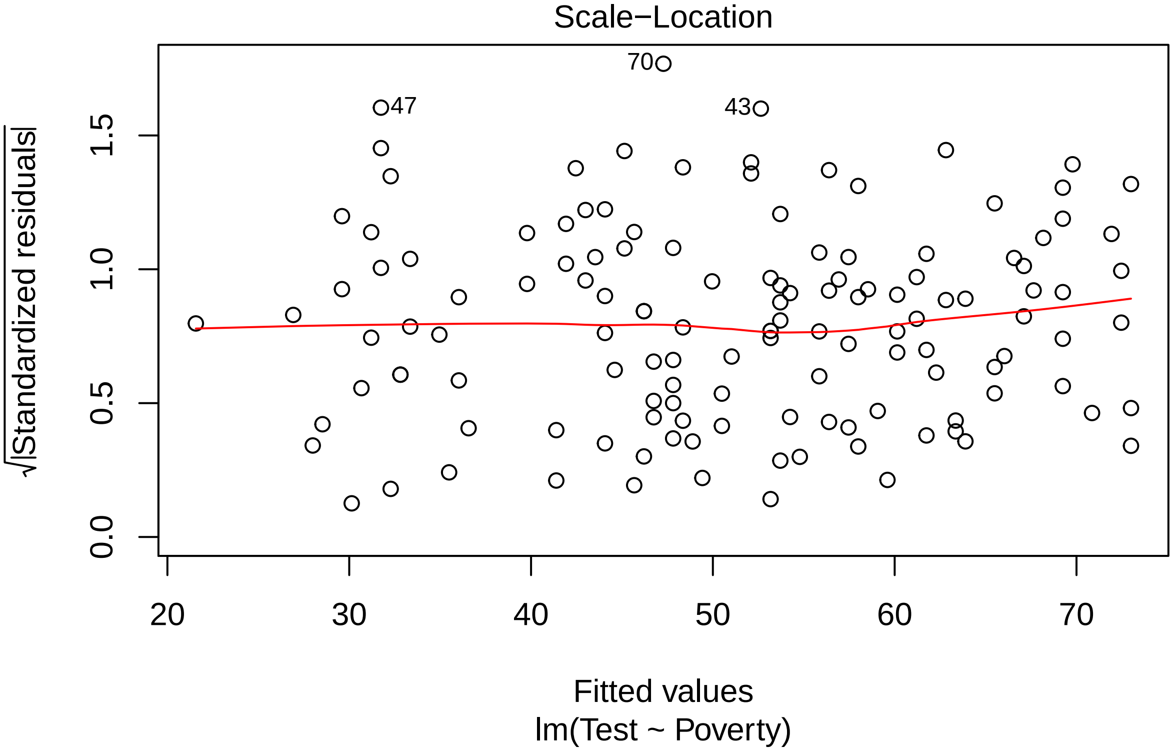
Question 3

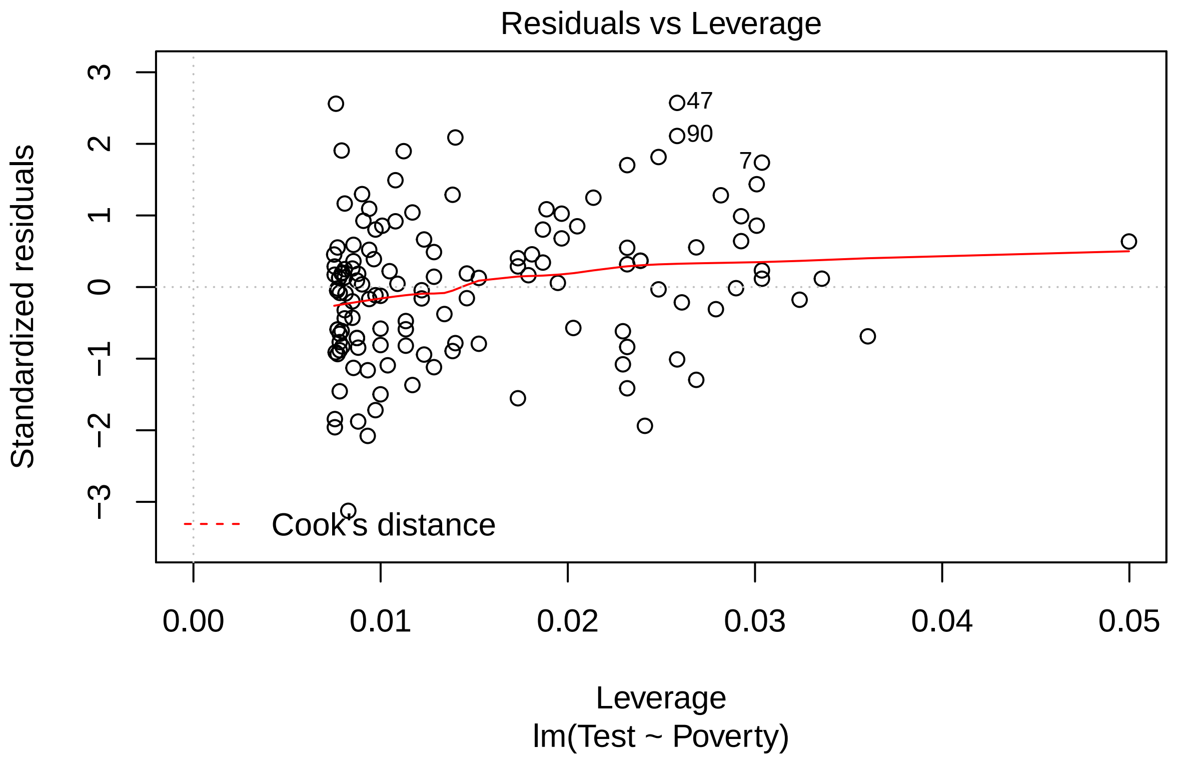
Describe any weaknesses in your model.

```
plot(iowa_lm)
```




Normal Q−Q
Standardized residuals
3
2
1
0
−1
−2
−3
43
47
70
−2
−1
0
1
2
Theoretical Quantiles
lm(Test ~ Poverty)


Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
70
47
43
20
30
40
50
60
70
Fitted values
lm(Test ~ Poverty)



The residual plot shows no clear pattern or "fanshape" indicating that a linear model is a good fit for this data and there is constant variance. The Normal Q-Q plot is relatively straight (except for some variation at the tails) which tells us that the data follows a normal distribution. There is no increasing or decreasing trend in the scale location plot which combined with the residual plot confirms that our data meets the constant variance condition. The residual vs. leverage plot shows no high leverage/influential points. Therefore, the model is relatively strong and only has weaknesses due to a slight curvature in the residual plot which potentially indicates that a different model may fit better. Likewise there is one point on the residual v. leverage plot that may indicate an outlier which could have skewed the model.

Question 4

What would you consider to be a well-performing school among schools with an 80% poverty rating?

```
predict(iowa_lm, data.frame(Poverty = 80), interval = "prediction", level = 0.95)
```

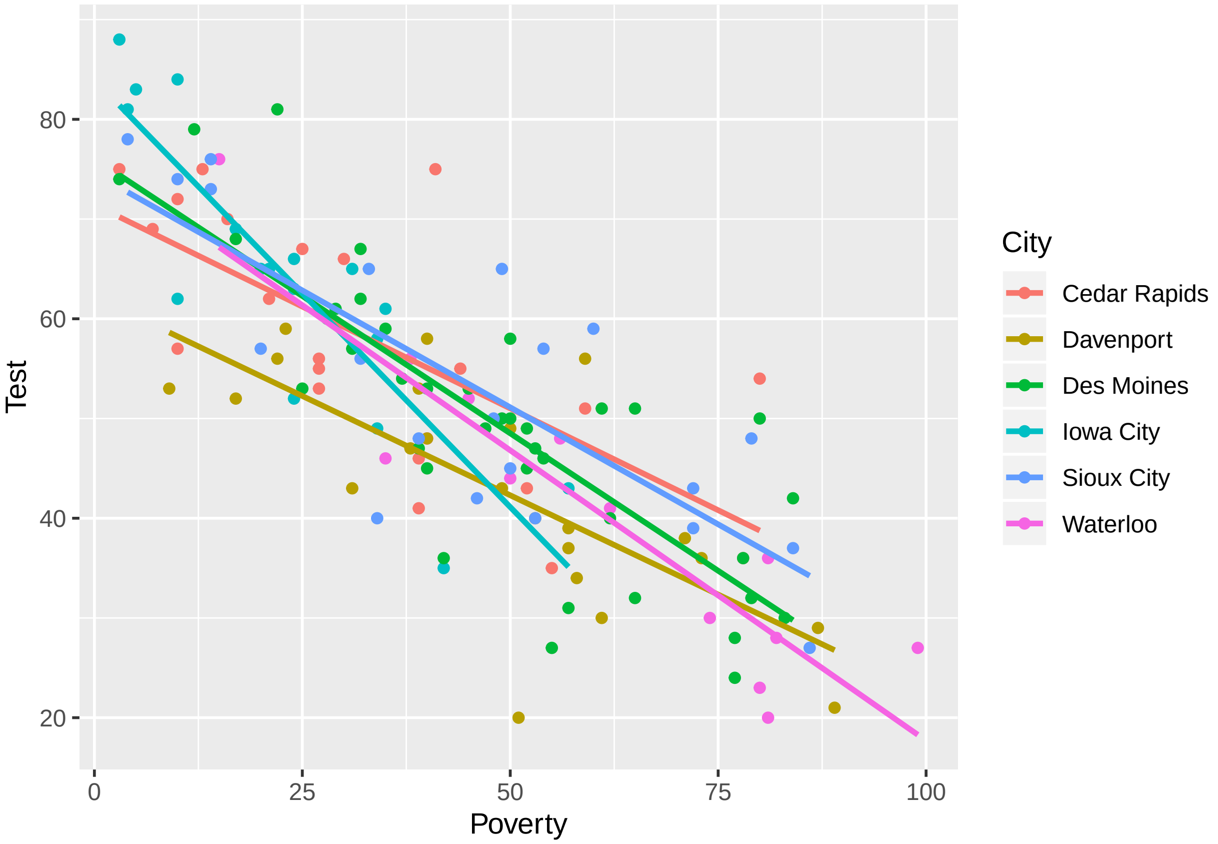
```
##        fit      lwr      upr
## 1 31.74375 14.17998 49.30752
```

A well-performing school among schools with an 80% poverty rating would recieve tests scores closer to the upper bound of the prediction interval, so that means a school with average test scores closer to ~49% can be considered well-performing when compared to the rest of its counterparts with an 80% poverty rating.

Question 5

Create a statistical graphic that illustrates how the relationship between poverty and test scores varies by city.

```
ggplot(iowatest, aes(Poverty, Test, colour = City)) +
  geom_point() +
  geom_smooth(se = FALSE, method = lm)
```



Question 6

What hypothesis test is the F-test in the summary output for Question 2 testing? State the hypotheses and the conclusion from the test.

The hypothesis test that the F-test is testing is comparing the null model to the full model. Specifically, the null hypothesis states that adding another variable (in this case poverty) and therefore a slope to the model does not improve the model's ability to explain variation in test scores. The alternative hypothesis says that adding the poverty variable (and resulting slope) improves the model's ability to explain the variation in test scores. Since the F-value is large (269.8) and p-value is small (less than 0.05), we reject the null hypothesis and thus see that adding another variable and therefore a slope improves our model's predictive ability as well as its ability to explain variation in test scores.