805518673MatthewSasakiHW5.rmd

Matthew Sasaki

4/29/2022

Question 1

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## 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
```

```
gender <- read.csv('armspans2022_gender.csv')
```

a

```
sum(gender$is.female)/nrow(gender)
```

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

0.348 of the class identified as female.

b

```
m1 <- lm(armspan~is.female, data=gender)
summary(m1)
```

```
##
## Call:
## lm(formula = armspan ~ is.female, data = gender)
##
## 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 69.76, which is the mean value when is.female = 0. This means the mean armspan for males was 69.76.

c

```
summary(m1)
```

```
##
## Call:
## lm(formula = armspan ~ is.female, data = gender)
##
## 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 slope is -7.73. which is the difference in average armspans for males and females.

d

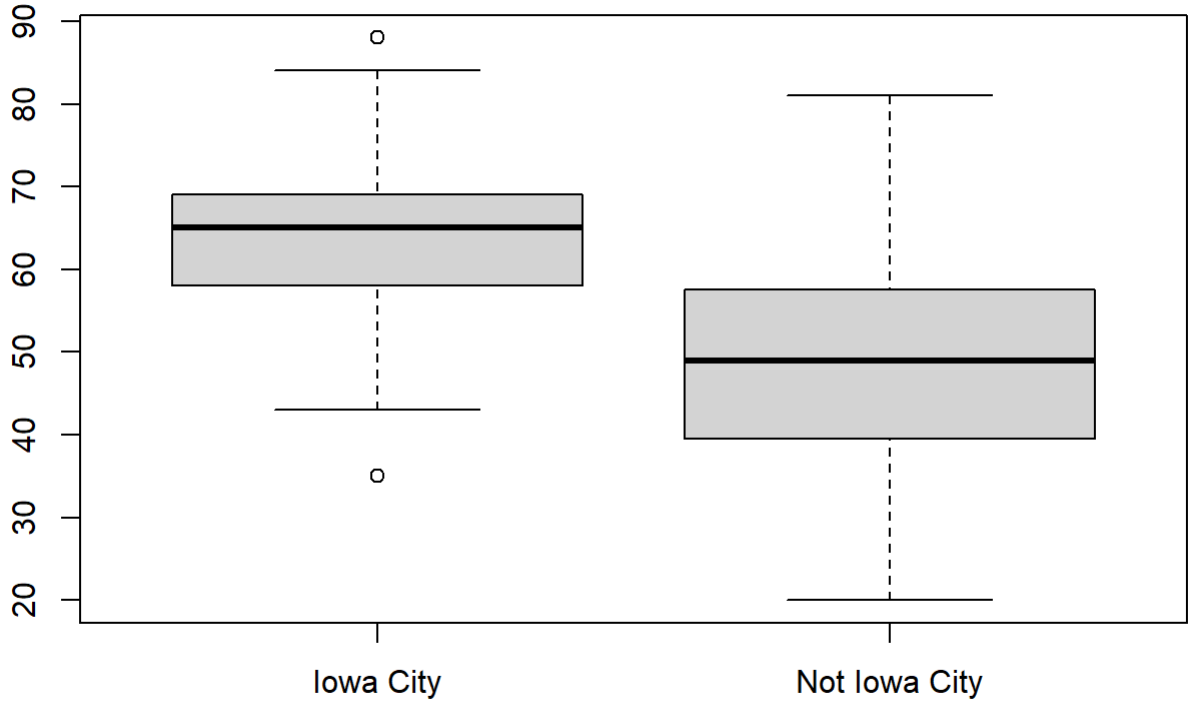
The test in question is whether or not there is a difference between armspans for males and females.

Question 2

```
iowa = read.table('iowatest.txt',sep = "\t", header=TRUE)
temp <- if_else(iowa$City=='Iowa City',1,0)
iowa$is_Iowa <- temp
is_Iowa_city <- iowa$is_Iowa == 1
t.test(iowa[is_Iowa_city,]$Test, iowa[!is_Iowa_city,]$Test, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  iowa[is_Iowa_city, ]$Test and iowa[!is_Iowa_city, ]$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
```

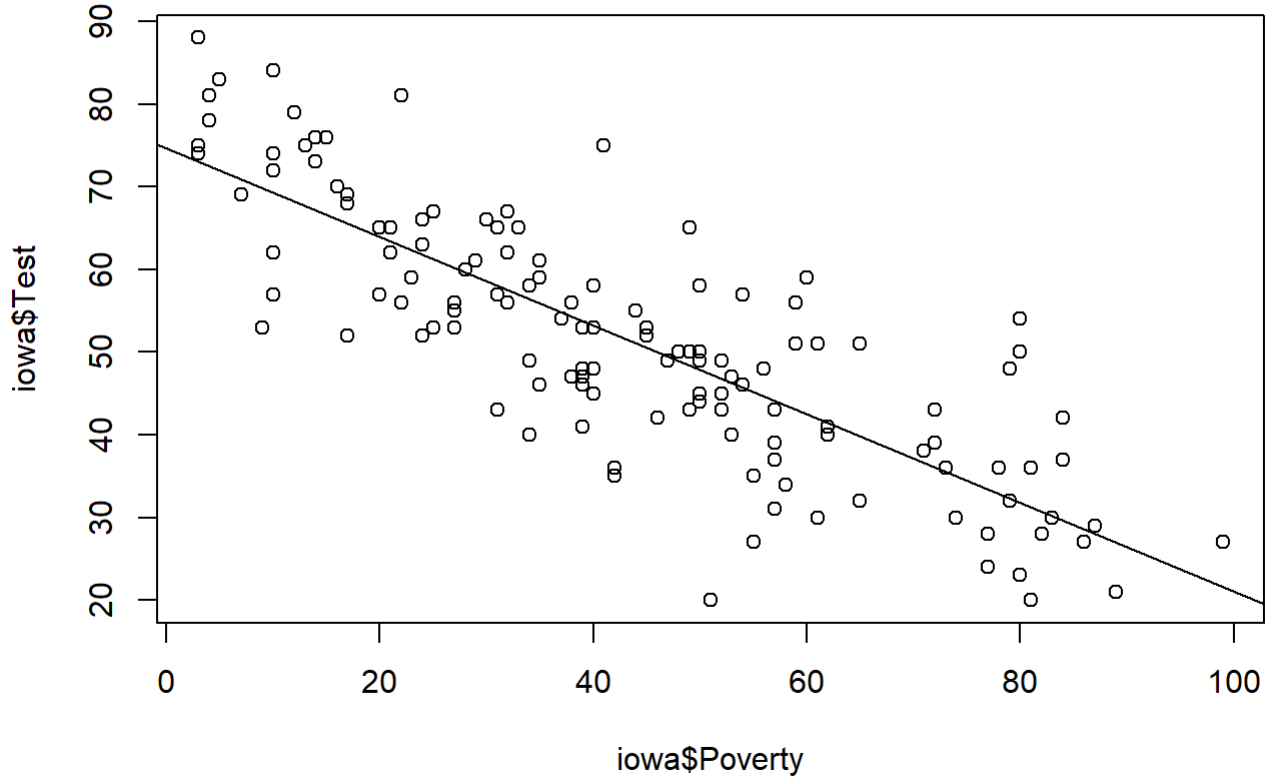
```
boxplot(iowa[is_Iowa_city,]$Test, iowa[!is_Iowa_city,]$Test, names = c("Iowa City","Not Iowa Cit
y"))
```



We see that the t-test yields a p value of 0.0004, which means we reject the null hypothesis and say that there is evidence supporting the hypothesis that students in Iowa city perform better than those not in Iowa city. The boxplot also shows a discrepancy in the scores.

Question 3

```
m2 <- lm(Test~Poverty, data=iowa)
plot(iowa$Poverty, iowa$Test)
abline(m2)
```



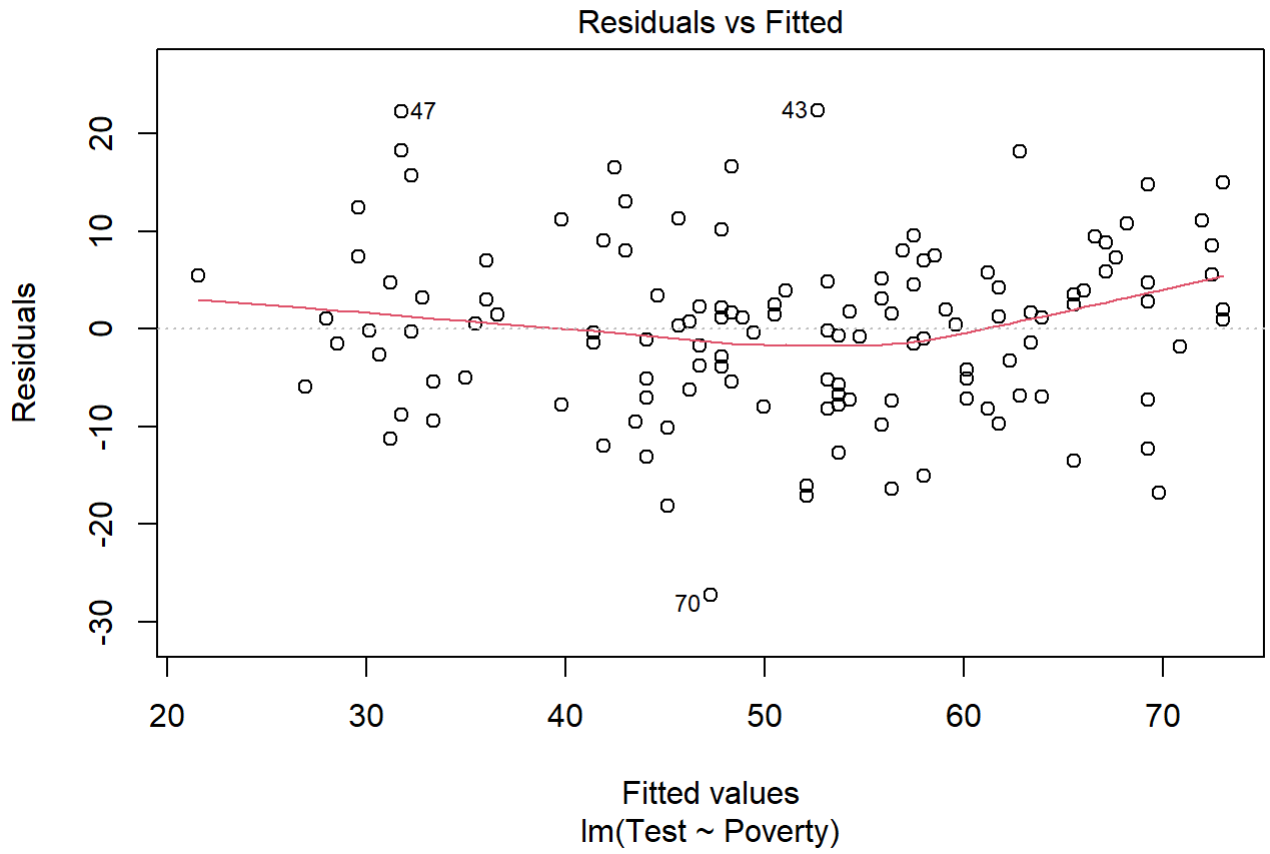

```
summary(m2)
```

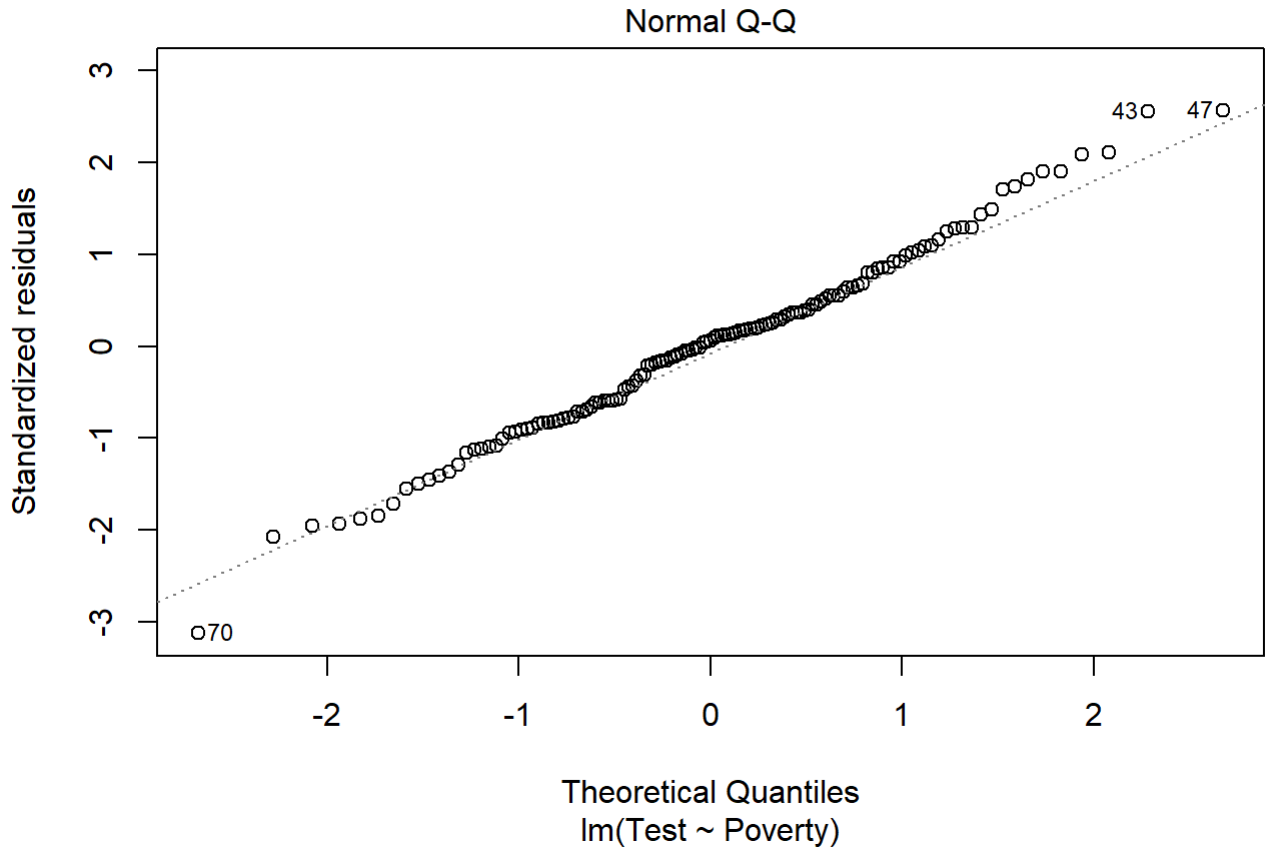
```
##
## Call:
## lm(formula = Test ~ Poverty, data = iowa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2812  -6.2097   0.5058   4.8252  22.3610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.60578    1.61325   46.25   <2e-16 ***
## 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
```

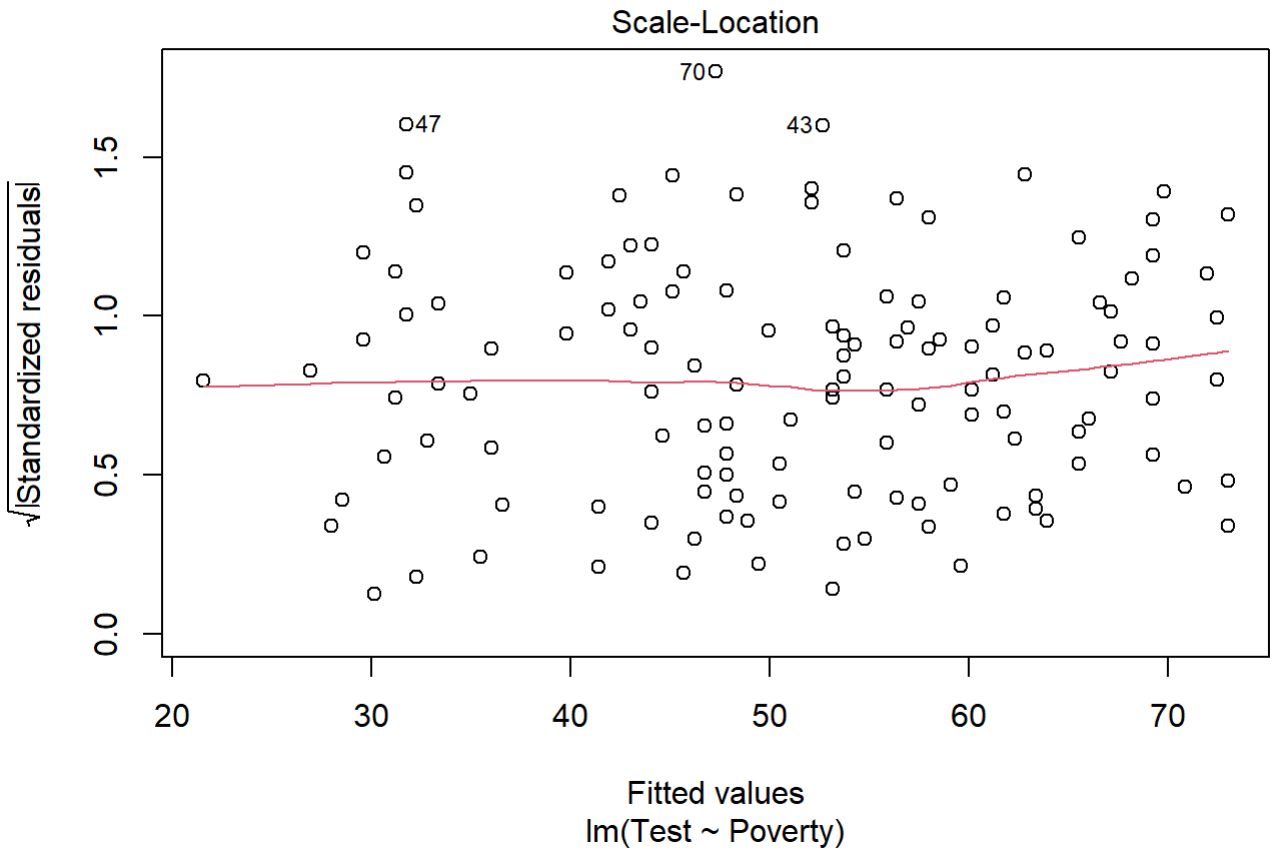
Based on this plot and an r^2 value of 0.67, we see that there is a least some correlation between poverty score and test score, with those having lower poverty scores tending to have higher test scores.

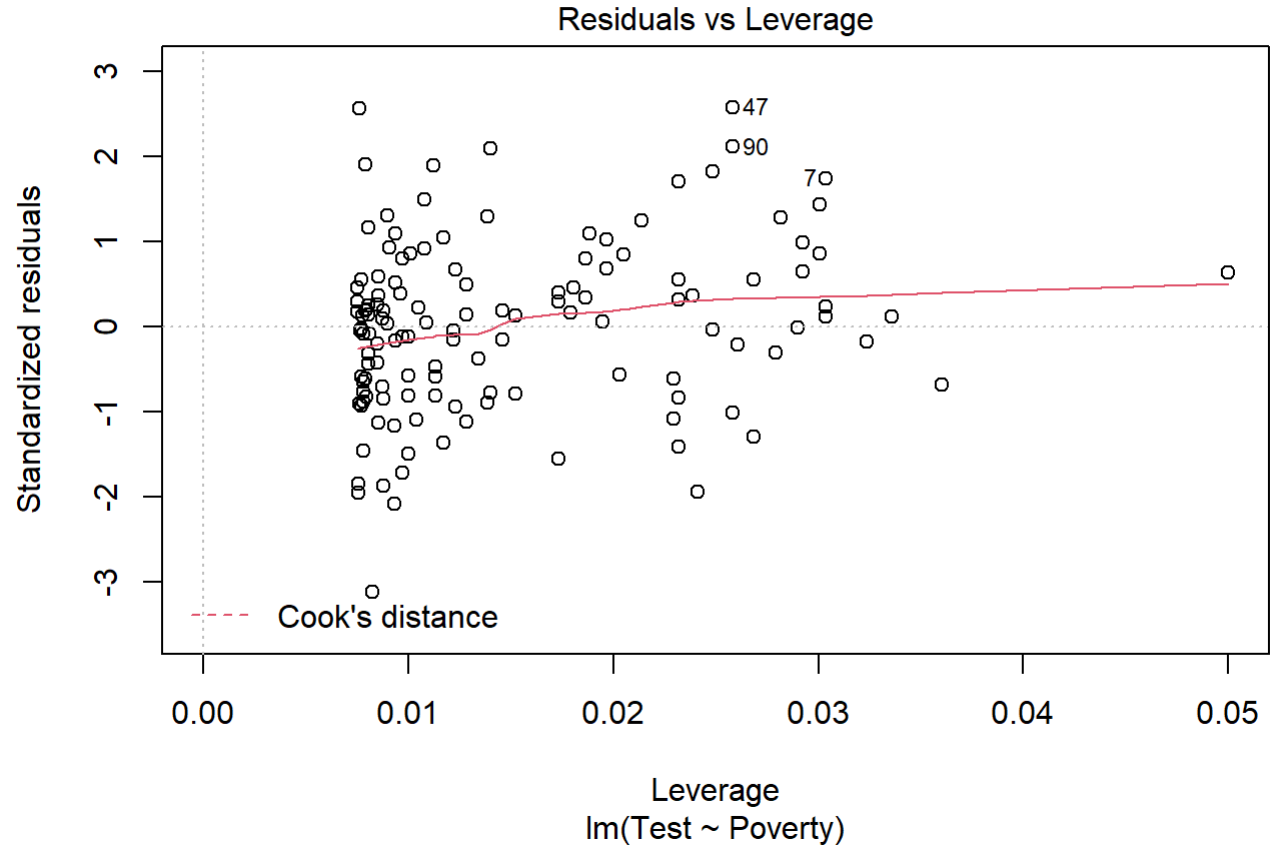
Question 4

```
plot(m2)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Test ~ Poverty)
47
43
70
20
10
0
-10
-20
-30
20
30
40
50
60
70


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


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


Residuals vs Leverage
47
90
7
Standardized residuals
-3
-2
-1
0
1
2
3
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
Leverage
lm(Test ~ Poverty)

For the residuals vs fitted plot, we don't really see a trend, with all the points being scatted around 0 relatively randomly. This indicates that the correlation is fairly linear. We also don't see higher variance at different x values, so the model has constant variance.

The normal QQ plot is a pretty straight line, so our data follows a normal distribution. Most of the points lie on the straight dotted line, so most of the points follow a normal distribution.

The scale location plot is similar to the earlier residual plot in that there doesn't really seem to be a trend. The standard residuals are similar for all x values, so this supports our assumption that our points have constant variance.

Question 5

```
leverage <- hatvalues(m2)
which.max(leverage)
```

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

```
bad <- sum(abs(rstandard(m2)) >2 & leverage>4/nrow(leverage))
bad
```

```
## [1] 0
```

The point with the highest leverage is in row 27, with a value of 0.05.

We see that we have no bad leverage points.

Question 6

```
summary(m2)
```

```
##
## 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 measures whether the model y = -0.536x + 74.6 is better than the simpler model y = 74.6. In other words, it is testing whether or not poverty affects their test score.