705604096_stats101a_hw5

Jade Gregory

2023-05-04

Question 0

a)

```
df_data <- 1
df <- 33
RSE <- 2.418
RSS <- RSE^2 * df
RSS
```

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

The RSS is 192.9419.

b)

```
f <- 87.17
SSreg <- f * RSS / df * df_data
SSreg
```

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

The SSreg is 509.6589.

c)

```
mean_SSreg <- SSreg / df_data
mean_SSreg
```

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

The mean SSreg is 509.6589.

d)

```
total_ss <- SSreg + RSS
total_ss
```

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

The total SS is 702.6008.

e)

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

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

The correlation coefficient is 0.994313.

Question 1

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

a)

```
new_arms <- arms[arms$is.female == 1,]
length(new_arms$is.female) / length(arms$is.female) * 100
```

```
## [1] 34.78261
```

34.78261% of the class identified as female.

b)

```
lm(armspan ~ is.female, data = arms)
```

```
##
## Call:
## lm(formula = armspan ~ is.female, data = arms)
##
## Coefficients:
## (Intercept)    is.female
##      69.759       -7.734
```

The regression equation would be 69.759 - 7.734 * (isfemale). The intercept means that when x = 0, y = 69.759. This is the expected value of Y, so this is the expected value of arm span for someone who identifies as male.

c) The mean tells us that there is a negative correlation between the two variables, arm span and is.female. When is.female = 0, that means the person identifies as male; when is.female is 1, they identify as female. Therefore, since there are two data points in this model, at x = 0 and x = 1, we can conclude that the negative correlation means that the average arm span for people identifying as female is 7.734 measurement units less than the average arm span for those identifying as male.

d)

```
summary(lm(armspan ~ is.female, data = arms))
```

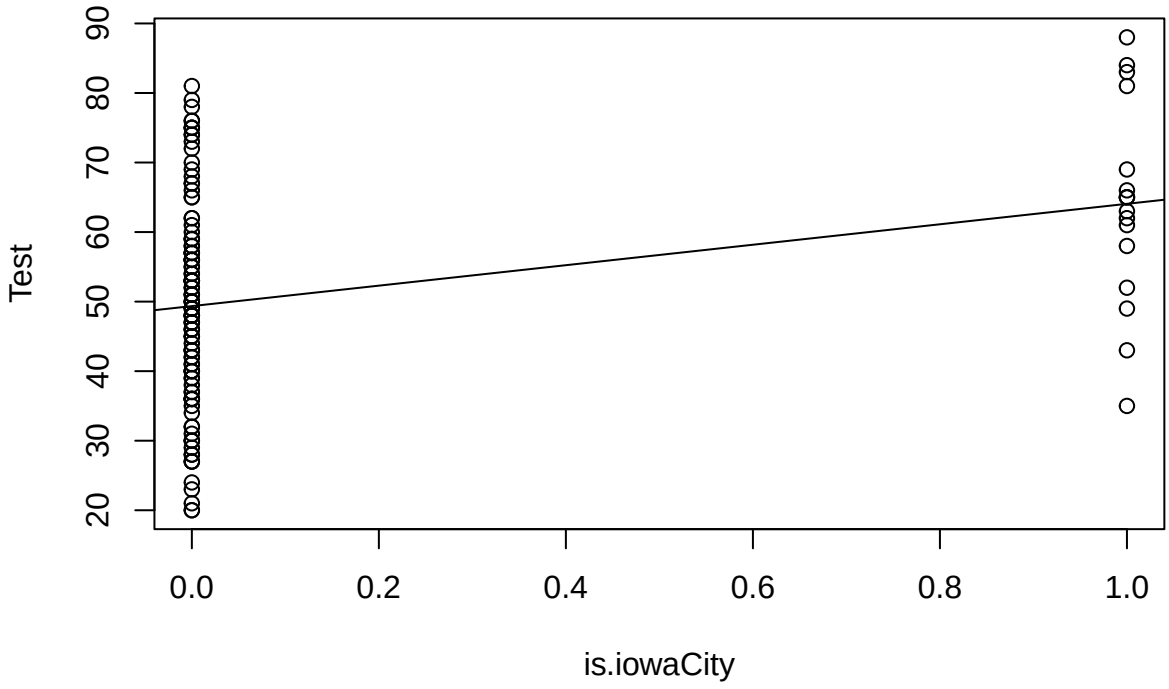
```
##
## Call:
## lm(formula = armspan ~ is.female, data = arms)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7586 -2.0248  0.2414  2.2414  8.2414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.7586     0.7399  94.284  < 2e-16 ***
## is.female    -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 null hypothesis states that the average arm span of males and females are the same while the alternative hypothesis states that the average arm span for males differs from the average arm span of females.

Question 2

```
iowa <- read.delim('iowatest.txt')
```

```
iowa$is.iowaCity <- iowa$City == 'Iowa City'
iowaCity.model <- lm(Test ~ is.iowaCity, data = iowa)
plot(Test ~ is.iowaCity, data = iowa)
abline(iowaCity.model)
```




```
summary(iowaCity.model)
```

```
##
## Call:
## lm(formula = Test ~ is.iowaCity, data = iowa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.353  -9.353  -0.353   7.647  31.647
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       49.353      1.347  36.626  < 2e-16 ***
## is.iowaCityTRUE   14.705      3.769   3.902 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.51 on 131 degrees of freedom
## Multiple R-squared:  0.1041, Adjusted R-squared:  0.09727
## F-statistic: 15.22 on 1 and 131 DF,  p-value: 0.000152
```

From our output of our linear model code, we can see that on average schools in Iowa City outperform schools outside of Iowa City by 14.705 on their test. In our hypothesis test performed, our null hypothesis claims that there is no difference between the average test scores and our alternative hypothesis states that there is a difference in test scores when comparing the two areas. Our p-value is 0.00152, which is lower than

our significance level of 0.05 so therefore we reject our null hypothesis and can conclude that on average there is a difference between the test scores of schools in Iowa City and outside of Iowa City. These both support our claim that Iowa City schools outperform the rest of the schools in different areas.

Question 3

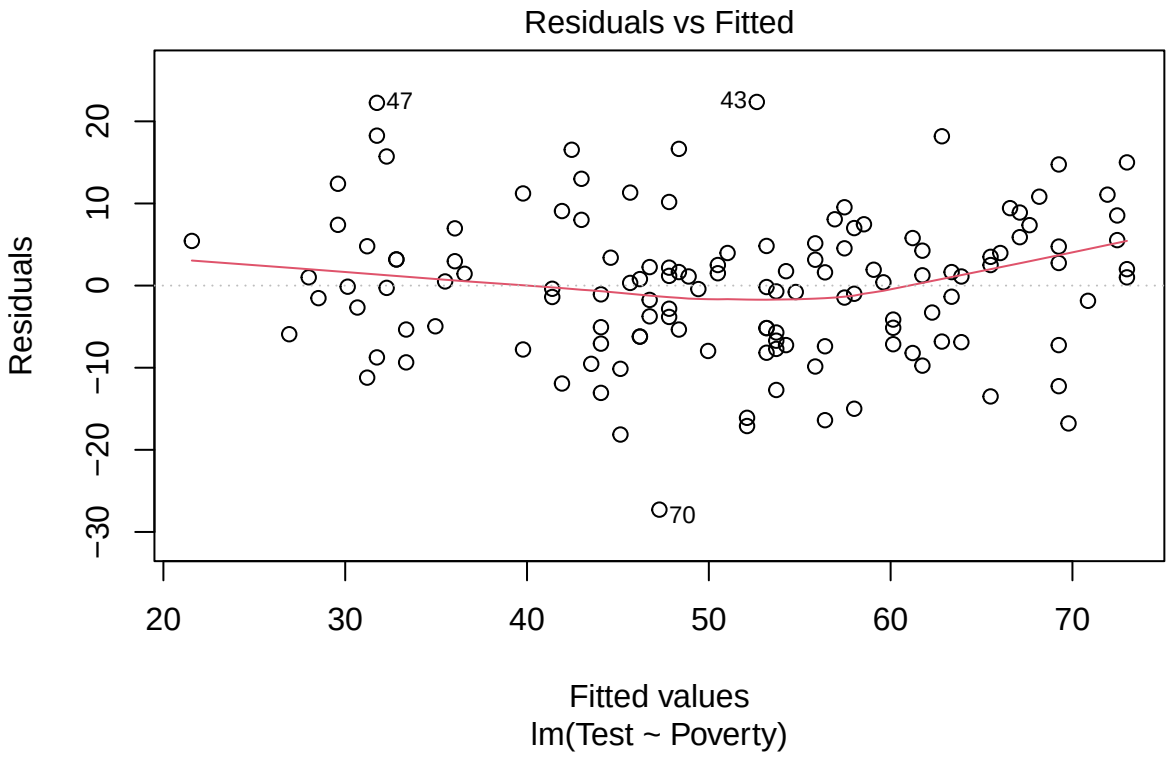
```
iowa.poverty <- lm(Test ~ Poverty, data = iowa)
summary(iowa.poverty)
```

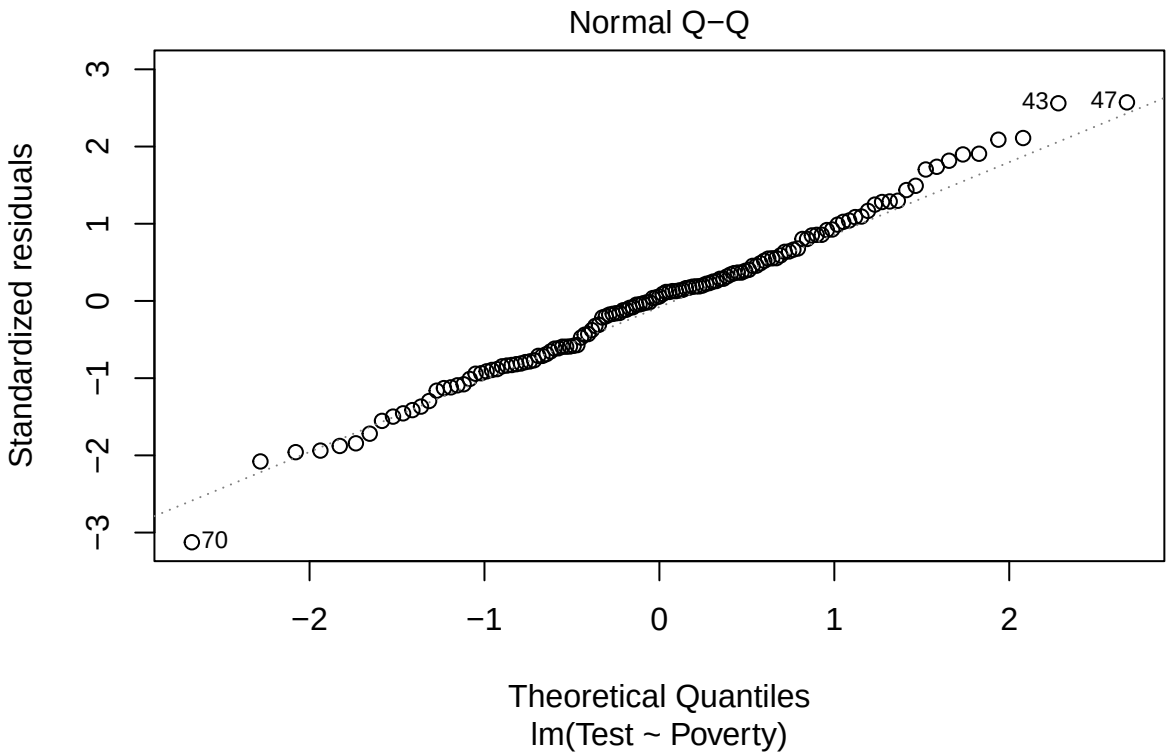
```
##
## Call:
## lm(formula = Test ~ Poverty, data = iowa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2812  -6.2097   0.5058   4.8252  22.3610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.60578    1.61325   46.25   <2e-16 ***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

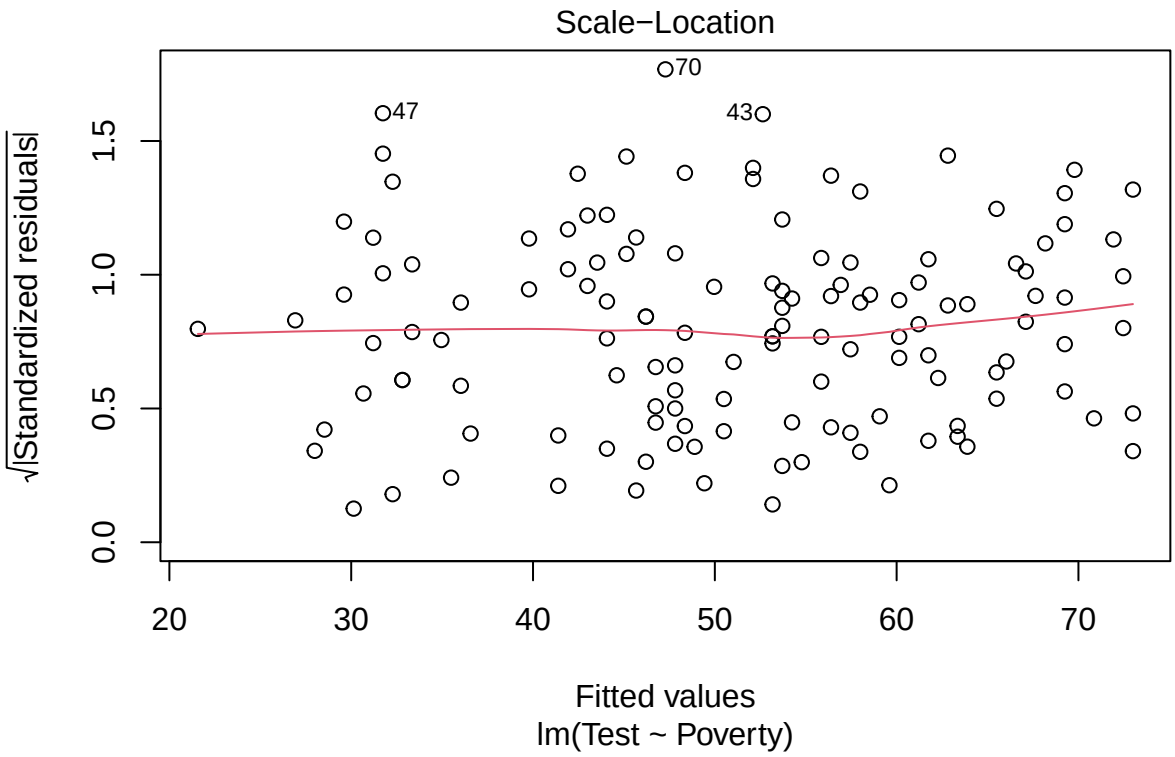
In our hypothesis test, our null hypothesis claims that poverty and test scores have no relationship while our alternative hypothesis claims that poverty and test scores do have a relationship. Our p-value is 2.2e-16, which is less than our significance level of 0.05, so therefore we can reject our null hypothesis and conclude that there is evidence that poverty is associated with test scores.

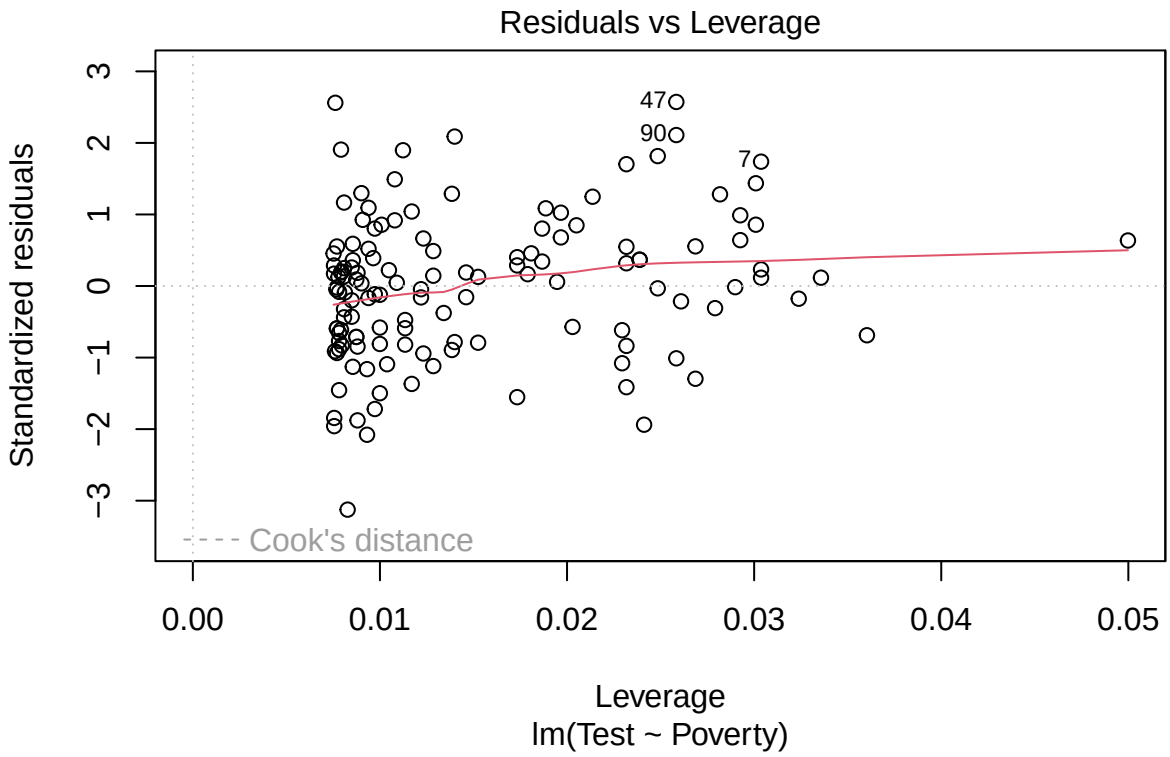
Question 4

```
plot(iowa.poverty)
```


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


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)



The first plot is the residuals vs fitted plot. There is no observable trend in our data points in this plot, suggesting that the linear model is a good fit for our data set. The second plot is the q-q plot. Since the data points in this graph follow the dashed line, there is evidence to assume that the data was drawn from a normal distribution. The third plot is the scale-location plot. Because the data points are plotted horizontally across the graph, we can assume there is constant variance in the model.

Question 5

```
iowa$leverage <- hatvalues(iowa.poverty)
iowa$rs <- rstandard(iowa.poverty)
max(iowa$leverage)
```

```
## [1] 0.04997855
```

```
iowa.highLeverage <- subset(iowa, leverage > 4 / nrow(iowa))
subset(iowa.highLeverage, abs(rs) > 2)
```

```
## [1] School      Poverty     Test        City        is.iowaCity leverage
## [7] rs
## <0 rows> (or 0-length row.names)
```

The row with the highest leverage point of 0.04997855 is row 27. We can conclude that there are no bad leverage points because none of them lie more than 2 standard residual deviations away from 0.

Question 6

In question 3, we are testing whether poverty and test scores have any relationship. The null hypothesis claims that poverty and test scores have no relationship while the alternative hypothesis states that test scores and poverty do have a relationship. Our p-value is 2.2e-16 which is less than our significance level of 0.05, so we can reject our null hypothesis and conclude that poverty and test scores are related.