Stats101A, Spring 2023 - Homework 3

Luke Villanueva - 206039397

05/05/23

Problem 0

a. RSS

$RSS = RSE^2 * df = 2.418^2 * 33 = 5.846724 * 33 =$
192.94

b. SSReg

$SSReg = F_{stat} * \frac{RSS}{df Res} df Reg = 87.17 * \frac{192.94}{33} * 1 = 87.17 * 5.8467 =$
509.6539

c. Mean SSReg

$MSSREG = \frac{SSReg}{df Reg} = \frac{509.6539}{1} =$
509.6539

d. Total Reg

$SYY = SSReg + RSS = 509.6539 + 192.94 =$
702.5939

e. R

$r = \sqrt{r^2} = \sqrt{0.7254} =$
0.8517

Problem 1

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
```

```
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
arms <- read.csv(file = paste0(getwd(),"/armspans2022_gender.csv"), header = TRUE)

glimpse(arms)
```

```
## Rows: 46
## Columns: 5
## $ height     <dbl> 74.00, 65.00, 60.00, 69.75, 70.00, 68.00, 64.00, 68.00, 68.~
## $ armspan    <dbl> 76.0, 65.0, 53.0, 69.0, 72.0, 70.5, 60.0, 67.0, 67.0, 60.0,~
## $ is.female  <int> 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1,~
## $ compmother <chr> "Taller", "Taller", "Shorter", "Taller", "Taller", "Taller"~
## $ compfather <chr> "Taller", "About the same", "Shorter", "About the same", "A~
```

a.

```
# proportion of females
mean(arms$is.female)
```

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

About 34.8% of the class reported to be female.

b.

```
# make linear model
m2 <- lm(armspan ~ is.female, data = arms)
m2$coefficients
```

```
## (Intercept)   is.female
##   69.758621   -7.733771
```

The intercept is 69.7586. This describes the expected value if is.female is null. In other words, 69.7586 is the mean value of non-females in the sample.

c.

The slope is -7.73. This describes the average difference between the armspans of female and non-female. In particular, this number says that on average, non-females have 7.73 less in armspan compared to females.

d.

The t-test is testing if the found slope from the sample can easily be obtained randomly through more samples. In other words,

Null: The true slope of the data is 0. Alternative: The true slope of the data is not 0.

Because the t-test is high, this means we can reject the null hypothesis and conclude that the true slope of the data is not 0, meaning there is some correlation between armspans and gender.

Problem 2

```
# text wrapping
library(stringi)


# load in
iowa <- read.delim(file = paste0(getwd(),"/iowatest.txt"), header = TRUE)

glimpse(iowa)
```

```
## Rows: 133
## Columns: 4
## $ School  <chr> "Coralville", "Hills", "Hoover", "Horn", "Kirkwood", "Lemme", ~
## $ Poverty <int> 20, 42, 10, 5, 34, 17, 3, 24, 21, 34, 24, 35, 4, 57, 24, 10, 3~
## $ Test    <int> 65, 35, 84, 83, 49, 69, 88, 63, 65, 58, 52, 61, 81, 43, 66, 62~
## $ City    <chr> "Iowa City", "Iowa City", "Iowa City", "Iowa City", "Iowa City~
```

```
colnames(iowa)
```

```
## [1] "School"  "Poverty" "Test"    "City"
```

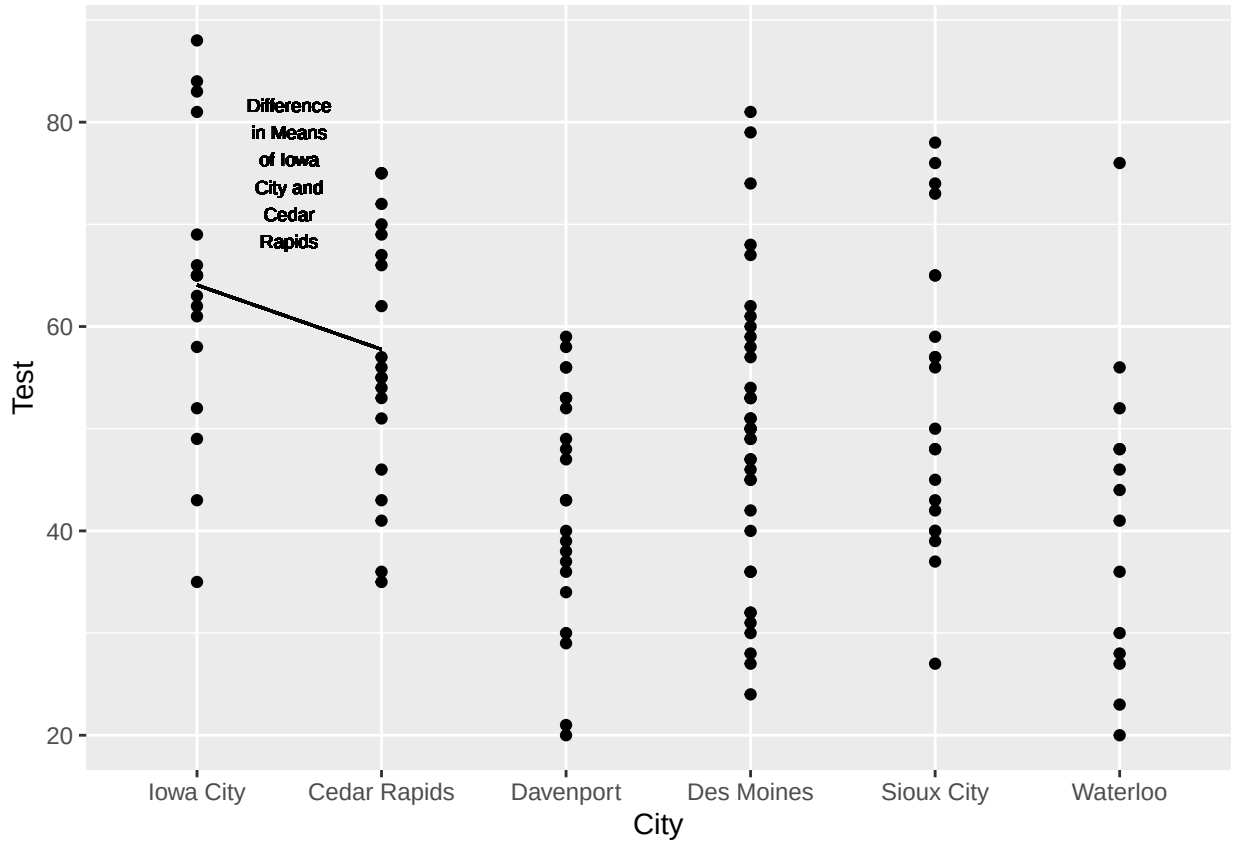
```
levels(factor(iowa$City))
```

```
## [1] "Cedar Rapids" "Davenport"    "Des Moines"   "Iowa City"    "Sioux City"
## [6] "Waterloo"
```

```
# change iowa city col into factor
iowa$City <- relevel(factor(iowa$City), "Iowa City")

# make linear model of test scores by cities in iowa
# change cities to be factor with Iowa City first
model <- lm(Test ~ City, iowa)

# get coefficients of model to calculate endpoints of segment
model$coefficients
```

```
##      (Intercept) CityCedar Rapids    CityDavenport   CityDes Moines
##        64.058824        -6.296919       -21.286096       -14.664087
##   CitySioux City     CityWaterloo
##       -10.773109       -22.987395
```

```r
# text wrapping function
wrapper <- function(x, ...) {paste(stri_wrap(x, ...), collapse = "\n")}

# plot
ggplot(iowa) + geom_point(aes(City, Test)) + geom_segment(aes(x = 1, xend = 2, y = 64.058824, yend = 57
```




```r
# get stats of model
summary(model)
```

```
##
## Call:
## lm(formula = Test ~ 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)        64.059      3.325  19.266  < 2e-16 ***
## CityCedar Rapids   -6.297      4.473  -1.408 0.161619
## CityDavenport     -21.286      4.427  -4.808 4.23e-06 ***
## CityDes Moines    -14.664      4.000  -3.666 0.000361 ***
## CitySioux City    -10.773      4.473  -2.409 0.017449 *
## CityWaterloo      -22.987      4.948  -4.646 8.34e-06 ***
```

```
## ---
## 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
```

Based on the statistics and graphics, we can see that generally, Iowa City outperforms all the other cities.

The summary dictates that the slope from Iowa City's mean to the other cities' is all negative, meaning that generally, all other cities' means are lower than Iowa City's.

However, based on the p-values of the slopes and on a 5% significance level for a two-tailed test, the difference between the means of Iowa City and Cedar Rapid is not statistically significant enough, for the p-value is only 0.1616. This implies that even if the means are different, the means could be different purely by chance, based on a 5% significance level.

Problem 3

```r
# make linear model for test score by poverty score
model <- lm(Test ~ Poverty, iowa)

# stats of model
summary(model)
```

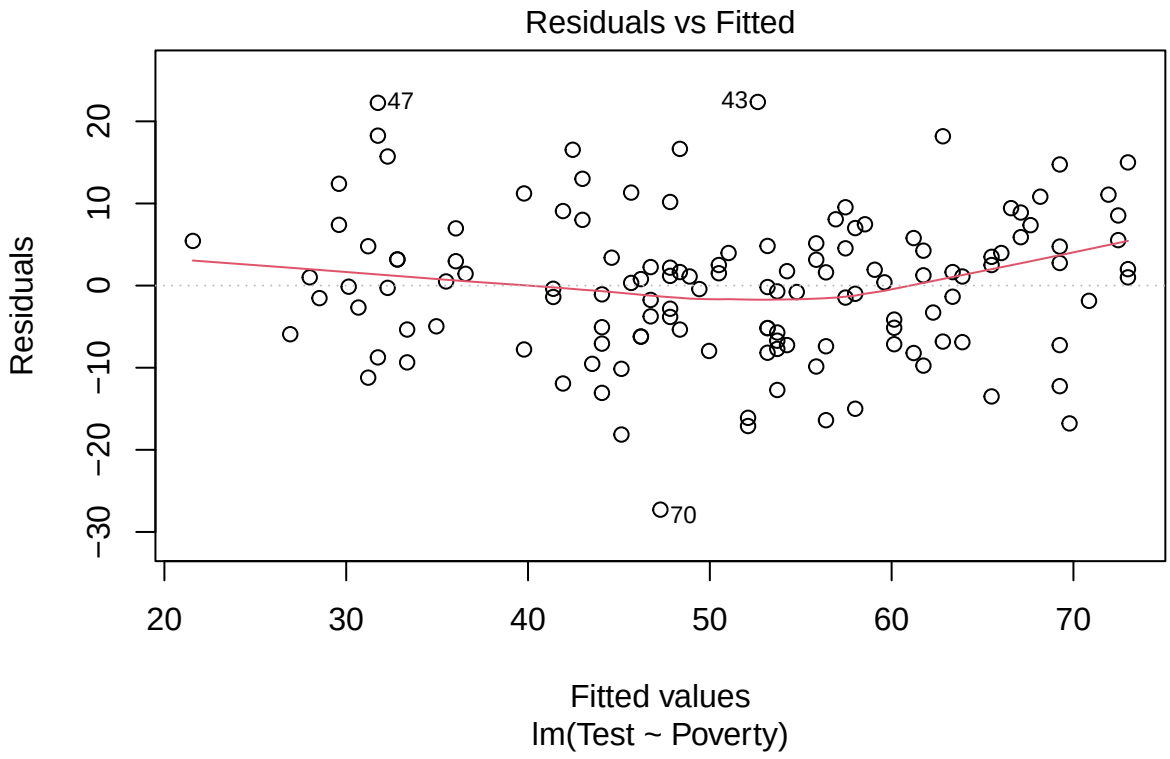
```
##
## Call:
## lm(formula = Test ~ Poverty, data = iowa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.2812  -6.2097   0.5058   4.8252  22.3610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.60578    1.61325   46.25   <2e-16 ***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

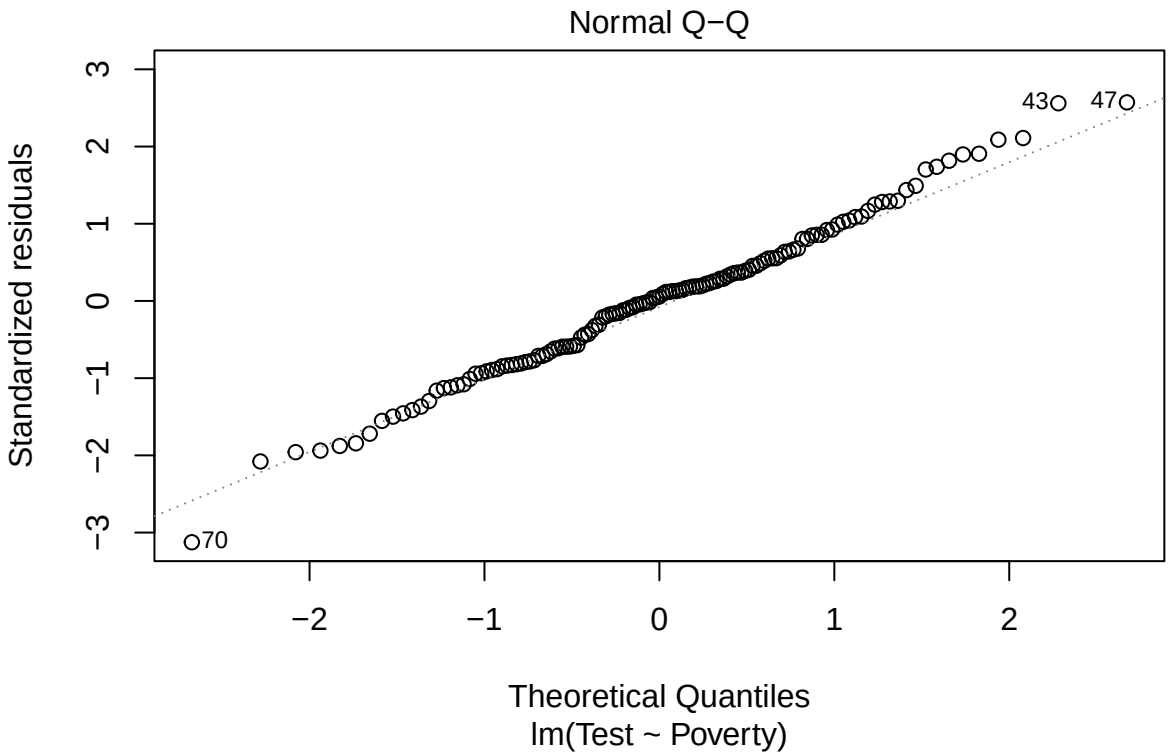
Based on the summary of the linear model's statistics, the slope is negative as Poverty increases. This implies the higher the Poverty score, the lower the expected test score will be. The p-values also support this claim, for the values are very close to 0. This implies the data's slope and intercepts are very unlikely to be obtained through random chance.

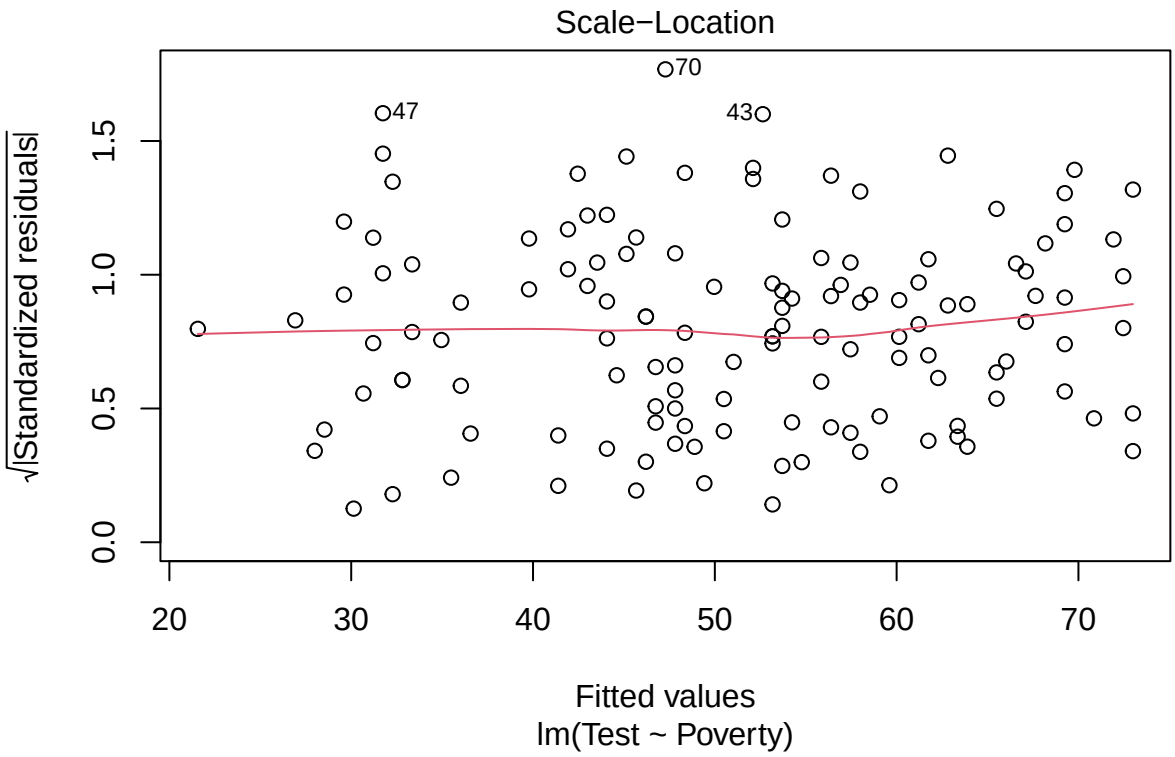
These two conclusions imply that there is indeed a correlation between Poverty score and Test score.

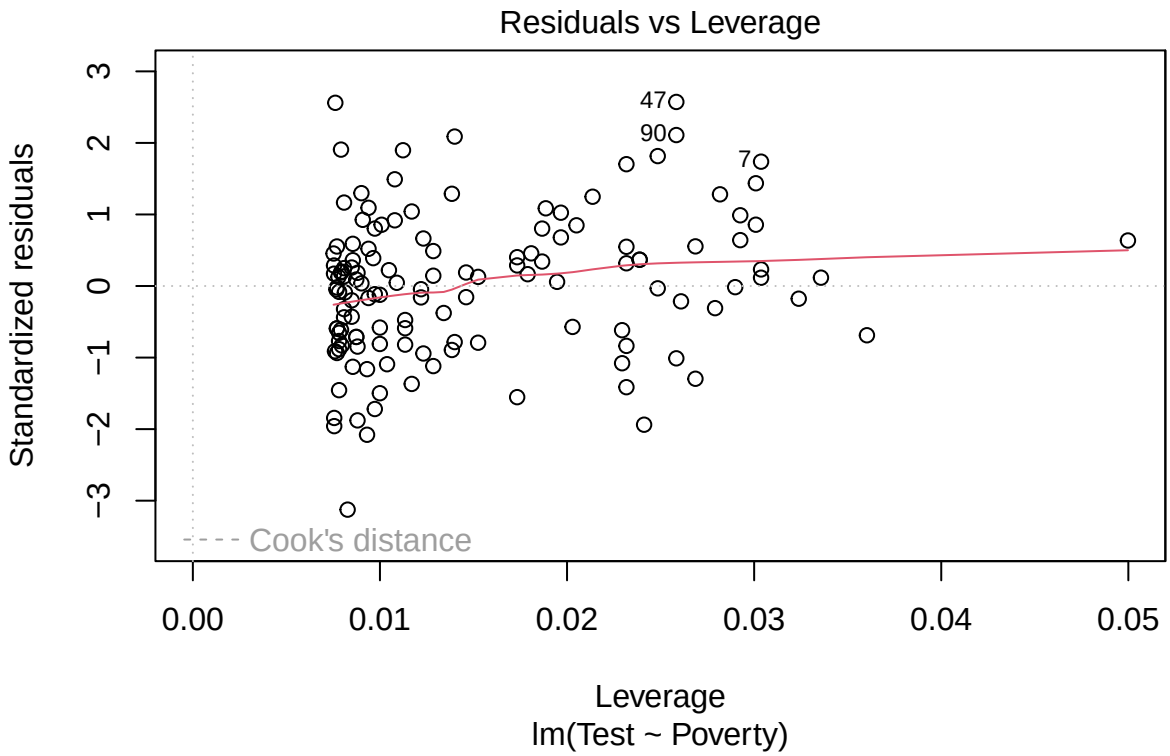
Problem 4

```
plot(model)
```




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)



Residuals: Based on the residual plot, it seems the residuals are randomly scattered and there seems to be no obvious trend in the graph. This implies that the data has strong linearity and predominantly consistent variance along the range of the predictor variables.

QQ: Based on the QQ plot, most of the values seem to be following a normal distribution. There are only a handful of values that seem to be outliers that do not follow the normal distribution. Thus, a general conclusion that normality can be assumed is a plausible deduction.

Scale-Location: The residuals on the scale-location plot seem to be randomly scattered and portray no obvious trend as the predictor variables changes. This implies that it can be assumed variance is constant.

Problem 5

```
# calculate hatvalues for every observation
hval <- hatvalues(model)

# get max hval index
hvalMax <- which(hval == max(hval))
# row of max hval
hvalMax
```

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

The observation that has the highest leverage is on row 27.

Regarding bad leverage points, based on Problem 4's leverage plot, we can use the guideline that a bad leverage point is more than 2 standard deviations away and the leverage amount is more than about 0.03 (because of the guideline formula 4/n for a high leverage amount).

```
# get standard residuals
rstan <- rstandard(model)
# get rstan greater than 2 std away
which(abs(rstan) > 2)
```

```
## 43 47 70 81 87 90
## 43 47 70 81 87 90
```

```
# get hvals greater than 4/n
which(hval > (4/nrow(iowa)))
```

```
##   7  27  46  64  67  89 109 120 126
##   7  27  46  64  67  89 109 120 126
```

```
# check if any observations match up
any(rstan %in% hval)
```

```
## [1] FALSE
```

```
any(hval %in% rstan)
```

```
## [1] FALSE
```

Based on these criteria, even if there are high leverage points and influential points, there are no points that fall into both categories of high leverage and high standard deviation. This implies there are no bad leverage points.

Problem 6

Null: The slope of all response variables (in this case, the Test score) is 0, with respect to the predictor variable, the Poverty score.

Alternative: The slope of all response variables (in this case, the Test score) is not 0, with respect to the predictor variable, the Poverty score.

Since the F-stat is 269.9 with a p-value close to 0, this implies that the slope obtained has a very unlikely chance of being obtained randomly. This means we can reject the null hypothesis. In conclusion, there is indeed a correlation between Poverty and Test scores, and the model fits well with the data.