Hw03ST430Yu

Haozhe (Jerry) Yu

2023-09-21

Question 1

```
educ <- as_tibble(read.table("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets
            #sep = "",
            strip.white=TRUE,
            col.name = c("Crime.Rate","High.School.Diploma")
            ))
educ
```

```
## # A tibble: 84 x 2
##    Crime.Rate High.School.Diploma
##         <int>               <int>
##  1       8487                  74
##  2       8179                  82
##  3       8362                  81
##  4       8220                  81
##  5       6246                  87
##  6       9100                  66
##  7       6561                  68
##  8       5873                  81
##  9       7993                  74
## 10       7932                  82
## # i 74 more rows
```

a. Find the least squares regression equation to predict the crime rate from the percent of individuals having at least a high school education. [Paste R or SAS output and then answer your question]

```
educm <- lm(Crime.Rate~High.School.Diploma,data=educ)

str(educm)
```

```
## List of 12
##  $ coefficients : Named num [1:2] 20518 -171
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "High.School.Diploma"
##  $ residuals    : Named num [1:84] 592 1649 1661 1519 568 ...
##   ..- attr(*, "names")= chr [1:84] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:84] -65175 9668 1528 1386 278 ...
##   ..- attr(*, "names")= chr [1:84] "(Intercept)" "High.School.Diploma" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:84] 7895 6530 6701 6701 5678 ...
##   ..- attr(*, "names")= chr [1:84] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:84, 1:2] -9.165 0.109 0.109 0.109 0.109 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:84] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "High.School.Diploma"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   ..$ qraux: num [1:2] 1.11 1.07
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 82
##  $ xlevels      : Named list()
##  $ call         : language lm(formula = Crime.Rate ~ High.School.Diploma, data = educ)
##  $ terms        :Classes 'terms', 'formula'  language Crime.Rate ~ High.School.Diploma
##   .. ..- attr(*, "variables")= language list(Crime.Rate, High.School.Diploma)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "Crime.Rate" "High.School.Diploma"
##   .. .. .. ..$ : chr "High.School.Diploma"
##   .. ..- attr(*, "term.labels")= chr "High.School.Diploma"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(Crime.Rate, High.School.Diploma)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "Crime.Rate" "High.School.Diploma"
##  $ model        :'data.frame':   84 obs. of  2 variables:
##   ..$ Crime.Rate         : int [1:84] 8487 8179 8362 8220 6246 9100 6561 5873 7993 7932 ...
##   ..$ High.School.Diploma: int [1:84] 74 82 81 81 87 66 68 81 74 82 ...
##   ..- attr(*, "terms")=Classes 'terms', 'formula'  language Crime.Rate ~ High.School.Diploma
##   .. .. ..- attr(*, "variables")= language list(Crime.Rate, High.School.Diploma)
##   .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. .. ..$ : chr [1:2] "Crime.Rate" "High.School.Diploma"
##   .. .. .. .. ..$ : chr "High.School.Diploma"
##   .. .. ..- attr(*, "term.labels")= chr "High.School.Diploma"
##   .. .. ..- attr(*, "order")= int 1
##   .. .. ..- attr(*, "intercept")= int 1
##   .. .. ..- attr(*, "response")= int 1
##   .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. .. ..- attr(*, "predvars")= language list(Crime.Rate, High.School.Diploma)
##   .. .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. .. ..- attr(*, "names")= chr [1:2] "Crime.Rate" "High.School.Diploma"
##  - attr(*, "class")= chr "lm"
```

The equation to predict crime rate (per 100,000 residents) from the percent of individuals in a country with at least a high school diploma is

Crime Rate = 2.05176×10^4 + -170.5751886High School Percent

b. Give the ANOVA Table for this regression analysis. [Paste R or SAS output]

c. Find SSE and MSE for this model.

d. What is the estimate of   from this analysis?

e. What percent of the variation in crime rates can be explained by the percent of high school graduates?

f. What is the correlation between crime rates and percent of high school graduates?

g. Based on your ANOVA table, is the linear relationship between X and Y statistically significant? Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

h. Give a scatter plot of crime rates vs. percent of high school graduates, with the regression line. Comment about linearity

i. Give the Residual Plot (residuals vs. fitted values). Test for Non-Linear and Non-constant variance.

j. Conduct Breusch-Pagan Test for the constancy of the error variance. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

k. Index Plot to test for Independence of errors.

l. Conduct Durbin-Watson Test. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value.

m. Outlier deduction test [Plot standardized Residuals versus fitted values]

n. Give a Histogram of the residuals and the density curve. Comment about the distribution of residuals.

o. Give a QQ-plot of the residuals to test for normality of error terms. Comment about the distribution of residuals.

p. Conduct a Shapiro-Wilk Test on the residuals. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value. Give the p-value for this test and explain what this means in terms of our model assumptions.

Question 2

a. Give a scatter plot

b. Find the least squares regression.

c. Give the Residual Plot (residuals vs. fitted values). Test for Non-Linear and Non-constant variance.

d. Conduct Breusch-Pagan Test for the constancy of the error variance.

e. Index Plot to test for Independence of errors.

c. Give the Residual Plot (residuals vs. fitted values). Test for Non-Linear and Non-constant variance.

d. Conduct Breusch-Pagan Test for the constancy of the error variance. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

e. Index Plot to test for Independence of errors.

f. Conduct Durbin-Watson Test. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value.

g. Outlier deduction test [Plot standardized Residuals versus fitted values]

h. Give a Histogram of the residuals and the density curve. Comment about the distribution of residuals.

i. Give a QQ-plot of the residuals to test for normality of error terms. Comment about the distribution of residuals.

j. Conduct a Shapiro-Wilk Test on the residuals. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value. Give the p-value for this test and explain what this means in terms of our model assumptions