Assignment: ASSIGNMENT 6

Name: Rattanavilay, Thip

Date: 2021-01-23

Graphical user interface, text, application

Description automatically generated

Output:

Text, letter

Description automatically generated

Chart, scatter chart

Description automatically generated

## Set the working directory to the root of your DSC 520 directory

> getwd()

[1] "/Users/trattanavilay/Documents/BU-Data\_SC\_Master/DSC520-Statistics-for-Data-Science/dsc520"

> setwd("/Users/trattanavilay/Documents/BU-Data\_SC\_Master/DSC520-Statistics-for-Data-Science/dsc520/")

> ## Load the `data/r4ds/heights.csv` to

> heights\_df <- read.csv("data/r4ds/heights.csv")

> ## Load the ggplot2 library

> library(ggplot2)

> ## Fit a linear model using the `age` variable as the predictor and `earn` as the outcome

> age\_lm <- lm(earn ~ age, data = heights\_df)

> ## View the summary of your model using `summary()`

> summary(age\_lm)

Call:

lm(formula = earn ~ age, data = heights\_df)

Residuals:

Min 1Q Median 3Q Max

-25098 -12622 -3667 6883 177579

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19041.53 1571.26 12.119 < 2e-16 \*\*\*

age 99.41 35.46 2.804 0.00514 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19420 on 1190 degrees of freedom

Multiple R-squared: 0.006561, Adjusted R-squared: 0.005727

F-statistic: 7.86 on 1 and 1190 DF, p-value: 0.005137

> ## Creating predictions using `predict()`

> age\_predict\_df <- data.frame(earn = predict(age\_lm, heights\_df), age=heights\_df$age)

> ## Plot the predictions against the original data

> ggplot(data = heights\_df, aes(y = earn, x = age)) +

+ geom\_point(color='blue') +

+ geom\_line(color='red', data = age\_predict\_df, aes(y = earn, x = age))

> mean\_earn <- mean(heights\_df$earn)

> ## Corrected Sum of Squares Total

> sst <- sum((mean\_earn - heights\_df$earn)^2)

> ## Corrected Sum of Squares for Model

> ssm <- sum((mean\_earn - age\_predict\_df$earn)^2)

> ## Residuals

> residuals <- heights\_df$earn - age\_predict\_df$earn

> ## Sum of Squares for Error

> sse <- sum(residuals^2)

> ## R Squared R^2 = SSM\SST

> r\_squared <- ssm / sst

> ## Number of observations

> n <- 1192 #From str(heights\_df)

> ## Number of regression parameters

> p <- 2

> ## Corrected Degrees of Freedom for Model (p-1)

> dfm <- (p - 1)

> ## Degrees of Freedom for Error (n-p)

> dfe <- (n - p)

> ## Corrected Degrees of Freedom Total: DFT = n - 1

> dft <- (n - 1)

> ## Mean of Squares for Model: MSM = SSM / DFM

> msm <- ssm / dfm

> ## Mean of Squares for Error: MSE = SSE / DFE

> mse <- sse / dfe

> ## Mean of Squares Total: MST = SST / DFT

> mst <- sst / dft

> ## F Statistic F = MSM/MSE

> f\_score <- msm / mse

> ## Adjusted R Squared R2 = 1 - (1 - R2)(n - 1) / (n - p)

> adjusted\_r\_squared <- 1 - (1 - r\_squared) \* (n - 1) / (n - p)

> ## Calculate the p-value from the F distribution

> p\_value <- pf(f\_score, dfm, dft, lower.tail=F)