# Assignment: ASSIGNMENT 6

# Name: Parker, Breanna

# Date: 2023-01-30

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

setwd("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1")

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

heights\_df <- read.csv("C:/Users/brean/OneDrive/Desktop/NucampFolder/projects/dsc520-1/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(heights\_df$age~heights\_df$earn)

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

summary(age\_lm)

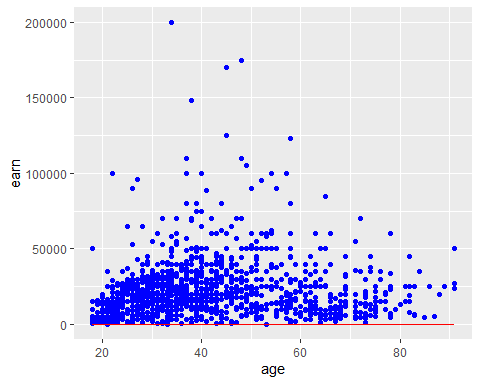
##   
## Call:  
## lm(formula = heights\_df$age ~ heights\_df$earn)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -25.150 -12.160 -3.840 8.712 49.566   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.985e+01 7.122e-01 55.954 < 2e-16 \*\*\*  
## heights\_df$earn 6.601e-05 2.354e-05 2.804 0.00514 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.82 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)  
mean\_earn

## [1] 23154.77

## Corrected Sum of Squares Total

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

## [1] 451591883937

## Corrected Sum of Squares for Model

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

## [1] 6.36801e+11

## Residuals

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

## Sum of Squares for Error

sse <- sum(residuals^2)  
sse

## [1] 1.088333e+12

## R Squared R^2 = SSM

r\_squared <- ssm/sst  
r\_squared

## [1] 1.410125

## Number of observations

n <- length(heights\_df$age)  
n

## [1] 1192

## Number of regression parameters

p <- 2

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

dfm <- p-1  
dfm

## [1] 1

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

dfe <- n - p  
dfe

## [1] 1190

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

dft <- n - 1  
dft

## [1] 1191

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

msm <- ssm/dfm  
msm

## [1] 6.36801e+11

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

mse <- sse/dfe  
mse

## [1] 914565780

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

mst <- sst/dft  
mst

## [1] 379170348

## F Statistic F = MSM/MSE

f\_score <- msm/mse  
f\_score

## [1] 696.2878

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

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

## [1] 1.41047

## Calculate the p-value from the F distribution

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

## [1] 3.339188e-121

names(summary(age\_lm))

## [1] "call" "terms" "residuals" "coefficients"   
## [5] "aliased" "sigma" "df" "r.squared"   
## [9] "adj.r.squared" "fstatistic" "cov.unscaled"

summary(age\_lm)$r.squared

## [1] 0.006561482

summary(age\_lm)$adj.r.squared

## [1] 0.005726659

summary(age\_lm)$fstatistic

## value numdf dendf   
## 7.859735 1.000000 1190.000000