assignment_08-09_MunjewarSheetal-02

Sheetal M

2023-02-12

Install and Load required packages :

```
# Package names
# packages <- c("ggplot2","dplyr","tidyr","magrittr","tidyverse","purrr")
packages <- c("ggplot2","dplyr","magrittr","tidyverse","purrr","pander","pandoc")

# Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Packages loading
invisible(lapply(packages, library, character.only = TRUE))
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8     v purrr   1.0.0
## v tidyr   1.2.1     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
## Set the working directory to the root of your DSC 520 directory
setwd("E:\\Data_Science_DSC510\\DSC520-Statistics\\dsc520")
```

```r
## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("data/r4ds/heights.csv")
# nrow(heights_df)
## Load the ggplot2 library
library(ggplot2)

# Fit a linear model
earn_lm <-  lm(heights_df$earn ~ heights_df$age + heights_df$height + heights_df$sex + heights_df$ed + h

# View the summary of your model
summary(earn_lm)
```

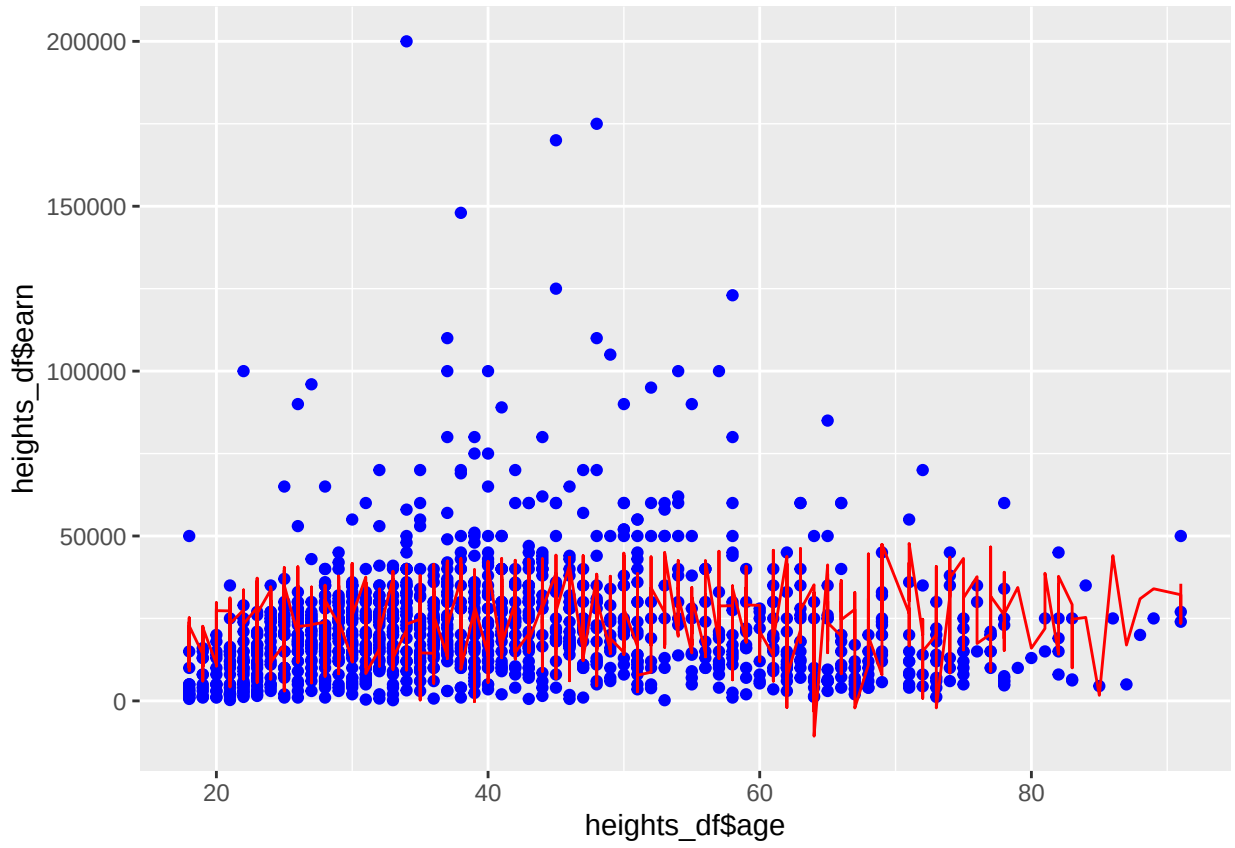
```
##
## Call:
## lm(formula = heights_df$earn ~ heights_df$age + heights_df$height +
##     heights_df$sex + heights_df$ed + heights_df$race, data = heights_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -39423  -9827  -2208   6157 158723
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -41478.4    12409.4  -3.342 0.000856 ***
## heights_df$age             178.3       32.2   5.537 3.78e-08 ***
## heights_df$height          202.5      185.6   1.091 0.275420
## heights_df$sexmale       10325.6     1424.5   7.249 7.57e-13 ***
## heights_df$ed             2768.4      209.9  13.190  < 2e-16 ***
## heights_df$racehispanic  -1414.3     2685.2  -0.527 0.598507
## heights_df$raceother       371.0     3837.0   0.097 0.922983
## heights_df$racewhite      2432.5     1723.9   1.411 0.158489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17250 on 1184 degrees of freedom
## Multiple R-squared:  0.2199, Adjusted R-squared:  0.2153
## F-statistic: 47.68 on 7 and 1184 DF,  p-value: < 2.2e-16
```

```r
predicted_df <- data.frame(
  earn = predict(earn_lm, heights_df),
  ed=heights_df$ed, race=heights_df$race, height=heights_df$height,
  age=heights_df$age, sex=heights_df$sex )

# - Reference https://www.youtube.com/watch?v=rfH7pCFvFT0 ( Linear Regression )
ggplot(data = heights_df, aes(x = heights_df$age, y = heights_df$earn)) +
  geom_point(color='blue') +
  geom_line(color='red',data = predicted_df, aes(x=predicted_df$age, y=predicted_df$earn))
```

```
## Warning: Use of `heights_df$age` is discouraged.
## i Use `age` instead.

## Warning: Use of `heights_df$earn` is discouraged.
## i Use `earn` instead.
```




```
## Compute deviation (i.e. residuals)
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 - predicted_df$earn)^2)

## Residuals
residuals <- heights_df$earn - predicted_df$earn

## Sum of Squares for Error
sse <- sum(residuals^2)

## R Squared
r_squared <- ssm/sst

## Number of observations
n <- 1192

## Number of regression paramaters
p <- 8

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

```r
## Degrees of Freedom for Error
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)
```