July 25, 2023

The results below are generated from an R script.

```
# Assignment: ASSIGNMENT 7
# Name: Tang, Xin
# Date: 2023-07-26

## Set the working directory to the root of your DSC 520 directory
setwd("~/dsc520")

## Load the `data/r4ds/heights.csv` to
library(ggplot2)

# Fit a linear model
earn_lm <-  lm(earn ~ ed + race + height + age + sex, data = heights_df)

# View the summary of your model
summary(earn_lm)

##
## Call:
## lm(formula = earn ~ ed + race + height + age + sex, 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 ***
## ed             2768.4      209.9  13.190  < 2e-16 ***
## racehispanic  -1414.3     2685.2  -0.527 0.598507
## raceother       371.0     3837.0   0.097 0.922983
## racewhite      2432.5     1723.9   1.411 0.158489
## height          202.5      185.6   1.091 0.275420
## age             178.3       32.2   5.537 3.78e-08 ***
## sexmale       10325.6     1424.5   7.249 7.57e-13 ***
## ---
## 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

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

```
  age=heights_df$age, sex=heights_df$sex
  )
View(predicted_df)

## 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 - 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_squared <- ssm/sst

## Number of observations
n <- nrow(heights_df)
## Number of regression paramaters
p <- 8
## Corrected Degrees of Freedom for Model
dfm <- p-1
## 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_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)
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19045)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United States.utf8  LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8 LC_NUMERIC=C
```

```
## [5] LC_TIME=English_United States.utf8
##
## time zone: America/Chicago
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.4.2
##
## loaded via a namespace (and not attached):
##  [1] vctrs_0.6.2      cli_3.6.1        knitr_1.43       rlang_1.1.1      xfun_0.39
##  [6] highr_0.10       generics_0.1.3   glue_1.6.2       labeling_0.4.2   colorspace_2.1-0
## [11] tinytex_0.45     scales_1.2.1     fansi_1.0.4      grid_4.3.1       munsell_0.5.0
## [16] evaluate_0.21    tibble_3.2.1     lifecycle_1.0.3  compiler_4.3.1   dplyr_1.1.2
## [21] pkgconfig_2.0.3  rstudioapi_0.14  farver_2.1.1     R6_2.5.1         tidyselect_1.2.0
## [26] utf8_1.2.3       pillar_1.9.0     magrittr_2.0.3   tools_4.3.1      withr_2.5.0
## [31] gtable_0.3.3
```

```
Sys.time()
```

```
## [1] "2023-07-25 21:17:20 CDT"
```