Assignment 7

Aarti Ramani

2023-02-10

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Masters/GitHub/Winter2022/Ramani-DSC520")

## Load the `data/r4ds/heights.csv` to
heights_df <- read.csv("data/r4ds/heights.csv")
names(heights_df)
```

```
## [1] "earn"   "height" "sex"    "ed"     "age"    "race"
```

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

```
## 
## Call:
## lm(formula = earn ~ ed + race + height + age + sex, data = heights_df)
## 
## Coefficients:
##  (Intercept)            ed  racehispanic     raceother     racewhite  
##     -41478.5        2768.4       -1414.3         371.0        2432.5  
##       height           age       sexmale  
##        202.5         178.3       10325.6
```

```
# 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, heights_df),
  ed=18, race='hispanic', height=71.7,
  age=35, sex='male'
)
nrow(predicted_df)
```

```
## [1] 1192
```

```
head(predicted_df)
```

```
##       earn ed     race height age  sex
## 1 38666.11 18 hispanic   71.7  35 male
## 2 28859.09 18 hispanic   71.7  35 male
## 3 23301.90 18 hispanic   71.7  35 male
## 4 32189.84 18 hispanic   71.7  35 male
## 5 27807.39 18 hispanic   71.7  35 male
## 6 20154.60 18 hispanic   71.7  35 male
```

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

```
## [1] 99302918657
## Residuals
residuals <- heights_df$earn - predicted_df$earn
length(residuals)
```

```
## [1] 1192
```

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

```
## [1] 3.52289e+11
```

```
## R Squared
r_squared <-  ssm/sst
r_squared
```

```
## [1] 0.2198953
```

```
## Number of observations
n <- nrow(heights_df)
n
```

```
## [1] 1192
```

```
## Number of regression paramaters
p <- 8
p
```

```
## [1] 8
```

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

```
## [1] 7
```

```
## Degrees of Freedom for Error
dfe <- n-p
dfe
```

```
## [1] 1184
```

```
## 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] 14186131237
```

```
## Mean of Squares for Error:   MSE = SSE / DFE
mse <- sse/dfe
mse
```

```
## [1] 297541356
```

```
## Mean of Squares Total:   MST = SST / DFT
mst <- sst/dft
mst
```

```
## [1] 379170348
```

```
## F Statistic
f_score <- msm/mse
f_score
```

```
## [1] 47.67785
```

```
## 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] 0.2152832
```