Assignment 07 Part 1- Regressions

David Pahmer

2022-05-15

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/users/pahme/onedrive/documents/github/dsc520")

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

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

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

```
## 
## Call:
## lm(formula = earn ~ age + ed + height + race + 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 ***
## age             178.3       32.2   5.537 3.78e-08 ***
## ed             2768.4      209.9  13.190  < 2e-16 ***
## height          202.5      185.6   1.091 0.275420    
## 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    
## 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 = data.frame(age=heights_df$age, ed=heights_df$ed, race=heights_df$rac
                                               height=heights_df$height, sex=heights_df$sex)),
  ed=heights_df$ed, race=heights_df$race, height=heights_df$height,
  age=heights_df$age, sex=heights_df$sex
  )
```

```r
## Compute deviation (i.e. residuals)
(mean_earn <- mean(heights_df$earn))
```

```
## [1] 23154.77
```

```r
## Corrected Sum of Squares Total
(sst <- sum((mean_earn - heights_df$earn)^2))
```

```
## [1] 451591883937
```

```r
## Corrected Sum of Squares for Model
(ssm <- sum((mean_earn - predicted_df$earn)^2))
```

```
## [1] 99302918657
```

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

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

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

```r
## R Squared
(r_squared <- ssm/sst)
```

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

```r
# Again this conforms to the value given in the regression calculation above

## Number of observations
(n <- length(predicted_df$age)) #or nrow(predicted_df)
```

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

```r
## Number of regression paramaters
(p <- 8)
```

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

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

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

```r
# I still don't get this yet.

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

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

```r
## Corrected Degrees of Freedom Total:   DFT = n - 1
(dft <- n-1)
```

```
## [1] 1191
```

```r
## Mean of Squares for Model:   MSM = SSM / DFM
(msm <- ssm / dfm)
```

```
## [1] 14186131237
```

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

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

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

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

```r
## F Statistic
(f_score <- msm / mse)
```

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

```r
## Adjusted R Squared R2 = 1 - (1 - R2)(n - 1) / (n - p)
(adjusted_r_squared <- 1-(1-r_squared)*(n-1)/(n-p))
```

```
## [1] 0.2152832
```