assignment_07_LastnameFirstname.R

12702

2020-10-17

```
# Assignment: ASSIGNMENT 7
# Name: kummarikunta, Vidyasagar
# Date: 2020-10-16

## Set the working directory to the root of your DSC 520 directory
setwd("/Users/12702/Desktop/MODatascience/DSC-520")

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

head(heights_df)
```

```
##    earn   height    sex ed age  race
## 1 50000 74.42444   male 16  45 white
## 2 60000 65.53754 female 16  58 white
## 3 30000 63.62920 female 16  29 white
## 4 50000 63.10856 female 16  91 other
## 5 51000 63.40248 female 17  39 white
## 6  9000 64.39951 female 15  26 white
```

```
# Fit a linear model
earn_lm <-  lm(formula = 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
```

```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
  )

## 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 <- 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)
```