Assignment 06 Part 0- Regressions

David Pahmer

2022-05-15

```r
## 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")

## Load the ggplot2 library
library(ggplot2)

## Fit a linear model using the `age` variable as the predictor and `earn` as the outcome
age_lm <-  lm(earn ~ age, data=heights_df)

## View the summary of your model using `summary()`
summary(age_lm)
```

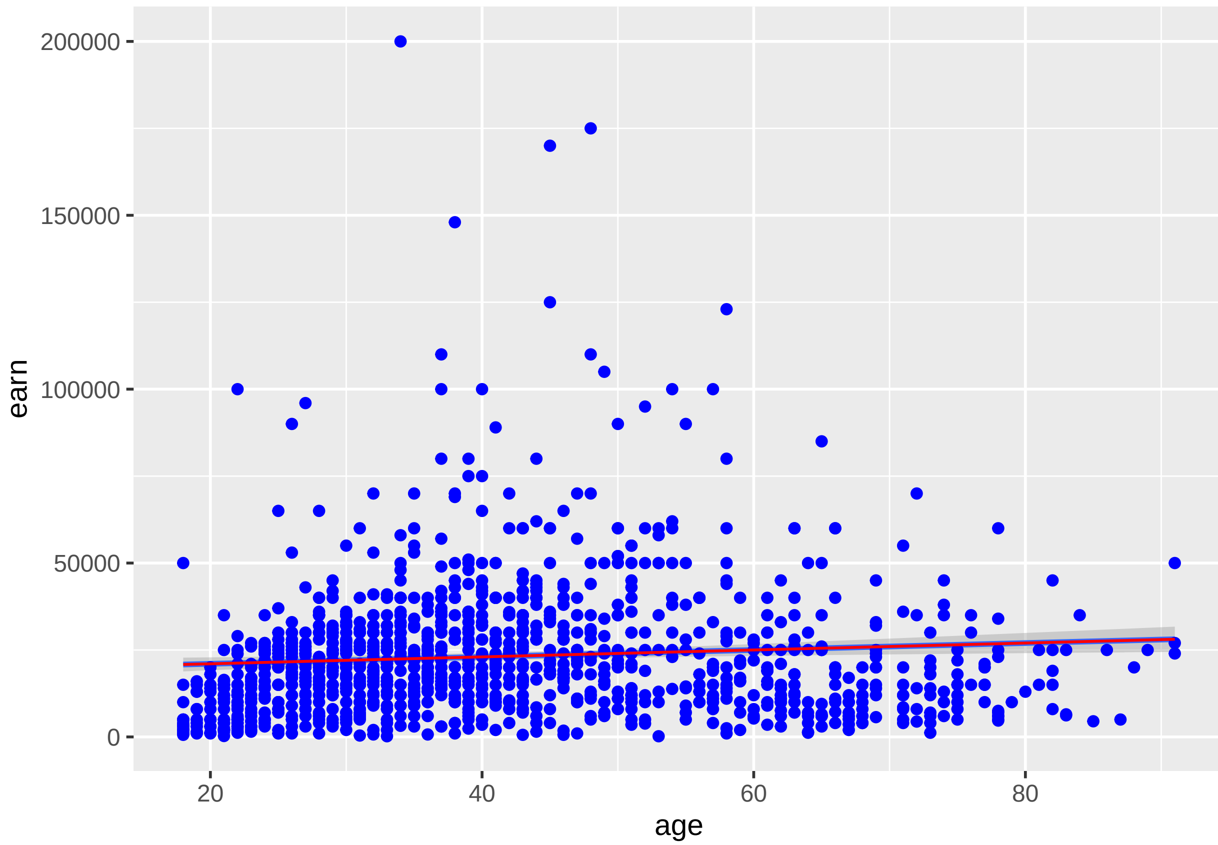
```
##
## Call:
## lm(formula = earn ~ age, data = heights_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -25098 -12622  -3667   6883 177579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19041.53    1571.26  12.119  < 2e-16 ***
## age            99.41      35.46   2.804  0.00514 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19420 on 1190 degrees of freedom
## Multiple R-squared:  0.006561,   Adjusted R-squared:  0.005727
## F-statistic:  7.86 on 1 and 1190 DF,  p-value: 0.005137
```

```r
## Creating predictions using `predict()`
age_predict_df <- data.frame(earn = predict(age_lm, newdata = data.frame(age=heights_df$age) ), age=hei

## Plot the predictions against the original data
ggplot(data = heights_df, aes(y = earn, x = age)) +
  geom_smooth(method="lm")+
  geom_point(color='blue') +

  geom_line(color='red',data = age_predict_df, aes(y=earn, x=age))
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
(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 - age_predict_df$earn)^2))
```

```
## [1] 2963111900
```

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

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

```
## [1] 448628772037
```

```r
## R Squared R^2 = SSM\SST
(r_squared <- ssm/sst)
```

```
## [1] 0.006561482
```

```r
# This matches the calculated R squared value shown in the summary of the regression above

## Number of observations
n <- length(heights_df$age)
## Number of regression parameters
p <- 2
## Corrected Degrees of Freedom for Model (p-1)
dfm <- p-1
# personally, I don't get this

## Degrees of Freedom for Error (n-p)
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)*(dft)/ (dfe)

## Calculate the p-value from the F distribution
p_value <- pf(f_score, dfm, dft, lower.tail=F)
```