Computer Proyect 1

Carmen Abans

2021-09-28

a. Use the read.cvs command to read the Earnings_and_Height.cvs data set into R. Use the attach command to attach the data set into R.

```
eah <- read.csv("Earnings_and_Height.csv")
attach(eah)
```

b. Print out an summary of the data set. In particular, find and report the sample average of the variables earnings, height and sex, respectively.

```
summary(eah)      # The summary shows: Min.    1st Qu.   Median    Mean    3rd Qu.    Max
```

```
##       sex              age              mrd              educ
##  Min.   :0.0000   Min.   :25.00   Min.   :1.000   Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:33.00   1st Qu.:1.000   1st Qu.:12.00
##  Median :0.0000   Median :40.00   Median :1.000   Median :13.00
##  Mean   :0.4419   Mean   :40.92   Mean   :2.362   Mean   :13.54
##  3rd Qu.:1.0000   3rd Qu.:48.00   3rd Qu.:4.000   3rd Qu.:16.00
##  Max.   :1.0000   Max.   :65.00   Max.   :6.000   Max.   :19.00
##     cworker          region           race          earnings
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 4726
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:23363
##  Median :1.000   Median :3.000   Median :1.000   Median :38925
##  Mean   :1.964   Mean   :2.551   Mean   :1.386   Mean   :46875
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.:84055
##  Max.   :6.000   Max.   :4.000   Max.   :4.000   Max.   :84055
##      height          weight        occupation
##  Min.   :48.00   Min.   : 80.0   Min.   : 1.000
##  1st Qu.:64.00   1st Qu.:140.0   1st Qu.: 2.000
##  Median :67.00   Median :163.0   Median : 5.000
##  Mean   :66.96   Mean   :170.4   Mean   : 6.011
##  3rd Qu.:70.00   3rd Qu.:190.0   3rd Qu.: 8.000
##  Max.   :84.00   Max.   :501.0   Max.   :15.000
```

```
summary(earnings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4726   23363   38925   46875   84055   84055
```

```
summary(height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   64.00   67.00   66.96   70.00   84.00
```

```
summary(sex)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4419  1.0000  1.0000
```

```
# The sex is a dummy variable where: 1=Male, 0 = Female
```

c. Run a regression of earnings on height. In particular, find and use a sentence to interpret the meaning of the regression coefficient of the variables height.

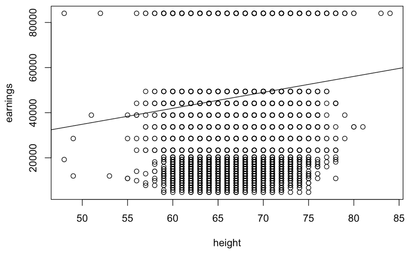
```
# Regression
ols <- lm(earnings ~ height)
summary(ols)
```

```
##
## Call:
## lm(formula = earnings ~ height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

```
# We have got that the height has this coefficients:

#   Estimate    707.67      (β1) This means that when the height increases by 1 (one inch taller)
#                           the earnings increase by $707.67.

#   Std. Error  50.49       (standard error of β1) This means that the average distance that the observed values
#                           deviate from the regression line is 50.49
#                           (The smaller the value, the closer our values are to the regression line)

#   t value     14.016      This is the coefficient divided by its standard error

#   Pr(>|t|)    <2e-16 ***  p-value
```

d. Plot a graph of earnings over height. e) On the graph, add a fitted line of the regression.

```
plot(earnings ~ height)
abline(ols)
```



f. Suppose Alex is 65 inches; Bob is 67 inches; Chris is 70 inches tall. Based on the regression, predict their corresponding earnings.

```
# Alex (Method 1)
-512.73 + 707.67*65
```

```
## [1] 45485.82
```

```
# Bob (Method 2)
ols$coefficient[1] + ols$coefficient[2] * 67
```

```
## (Intercept)
##    46901.26
```

```
# Chris(Method 3)
predict(ols, data.frame(height=70))
```

```
##        1
## 49024.28
```

g. Find the R2 and SER from the regression in part (c). Use a sentence to interpret each of them.

```
#     When we did the summary of the ols we got a residual standard error (SER) of $26780 and it is the measure
#     of the spread of the error term u (in this case is quite large so is not a good thing).

#     We also got the R2 which it appears to be 0.01088. This means that approximately 1.09% of earnings are
#     explained by the height.
```

h. Based on the regression in part (c), find the p-value of the variables height and perform a t-test.

```
#     The summary of the ols shows that our p-value is smaller than 2.2e-16.
#     If we do the t-test based on the p-value we get that the absolute value of the p-value is smaller than 1.96
#     For that reason we reject the null (H0: β1=0).
#     That means that β1!=0 so there is a relationship between height and earnings.
```

i. Based on the regression in part (c), use the confint command to calculate the Confidence Interval (CI) of the variables height. Does your CI give you the same t-test conclusion?

```
confint(ols)
```

```
##                 2.5 %     97.5 %
## (Intercept) -7151.2994 6125.8322
## height        608.7078  806.6353
```

```
# We've got that the CI = [608.7078, 806.6353]. Since 0 is out of the CI we also end up rejecting the null.
# This is to be expected as all three t-test methods are equivalent.
```

j. Run a regression of earnings on sex. For both the regression intercept and coefficient of the variables sex, use a sentence to interpret its meaning.

```
ols2<- lm(earnings ~ sex)
summary(ols2)
```

```
##
## Call:
## lm(formula = earnings ~ sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -43733 -22258  -6696  35595  38434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45621.0      269.2 169.455  < 2e-16 ***
## sex           2838.8      405.0   7.009 2.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26890 on 17868 degrees of freedom
## Multiple R-squared:  0.002742,   Adjusted R-squared:  0.002686
## F-statistic: 49.13 on 1 and 17868 DF,  p-value: 2.485e-12
```

```
# We've got that Earnings = 45621 + 2838.8 x Male (as the sex variables are 1=Male, 0 = Female).
# This means that men earn on average $2838.8 more than women.
# We can also see that the mean earnings of women is $45621.
# And if we want to know the mean earnings of men we just need to set the sex = 1 which results in:
45621.0+2838.8
```

```
## [1] 48459.8
```