Empirical Exercise - E4.2

Chi-Yuan Fang

2021-03-21

On the text website, http://www.pearsonglobaleditions.com, you will find the data file **Earnings__and__Height**, which contains data on earnings, height, and other characteristics of a random sample of U.S. workers. A detailed description is given in **Earnings__and__Height__Description**, also available on the website. In this exercise, you will investigate the relationship between earnings and height.

a. What is the median value of height in the sample?

Solution

```
# import data
library(readxl)
Earnings_and_Height <- read_xlsx("Earnings_and_Height/Earnings_and_Height.xlsx")

median(Earnings_and_Height$height)
```

```
## [1] 67
```

b. i. Estimate average earnings for workers whose height is at most 67 inches.
   ii. Estimate average earnings for workers whose height is greater than 67 inches.
   iii. On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

Solution

```
# create "group" variable
group <- c()

for (i in 1:length(Earnings_and_Height$height)){
  if (Earnings_and_Height$height[i] <= 67){
    # group 0: height <= 67
    group[i] <- c(0)
  } else {
    # group 1: height > 67
    group[i] <- c(1)
  }
}

Earnings_and_Height <- cbind(Earnings_and_Height, group)

E42b <- function(x){
  # sample mean
  mu <- mean(x)

  # sample standard deviation (standard error)
  se <- sd(x)/sqrt(length(x))
```

```r
  # test
  test <- t.test(x,
                 alternative = c("two.sided"),
                 mu = 0, # H0
                 conf.level = 0.95) # alpha = 0.05

  # 95% confidence interval
  lower <- round(test$conf.int[1], digit = 4)
  upper <- round(test$conf.int[2], digit = 4)
  CI <- paste(lower, "-"  ,upper)

  Table <- data.frame(mu, se, CI)
  colnames(Table) <- c("Mean", "Standard Error", "95% Confidence Interval")

  Table

}

# i. # ii.
tapply(Earnings_and_Height$earnings, Earnings_and_Height$group, E42b)
```

```
## $`0`
##       Mean Standard Error 95% Confidence Interval
## 1 44488.44       265.4948 43968.0133 - 45008.8585
##
## $`1`
##       Mean Standard Error 95% Confidence Interval
## 1 49987.88       305.4062 49389.1973 - 50586.5544
```

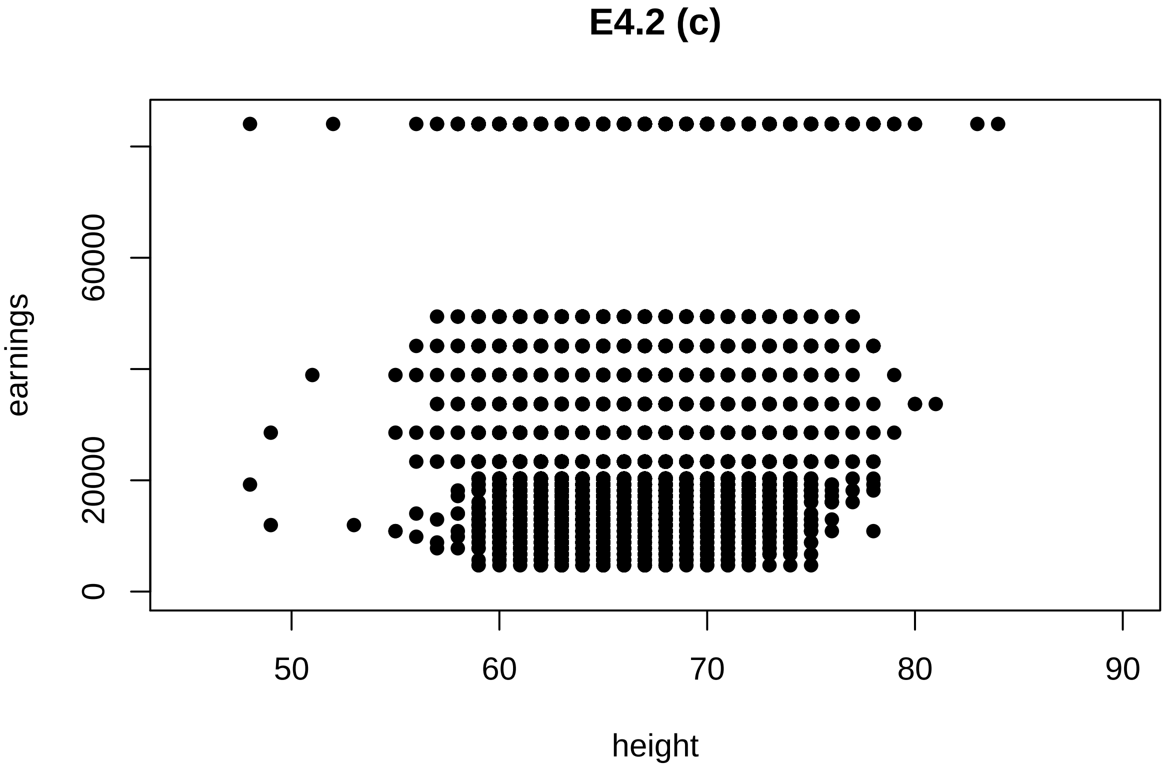
```r
# height <= 67
Earnings_and_Height_i <- Earnings_and_Height[Earnings_and_Height$height <= 67, ]

# height > 67
Earnings_and_Height_ii <- Earnings_and_Height[Earnings_and_Height$height > 67, ]

# iii. 95% CI for difference
t.test(Earnings_and_Height_ii$earnings, Earnings_and_Height_i$earnings,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95) # alpha = 0.05
```

```
##
##  Welch Two Sample t-test
##
## data:  Earnings_and_Height_ii$earnings and Earnings_and_Height_i$earnings
## t = 13.59, df = 16624, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4706.237 6292.643
## sample estimates:
## mean of x mean of y
##  49987.88  44488.44
```

c. Construct a scatterplot of annual earnings (*Earnings*) on height (*Height*). Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of *Earnings*). Why? (Hint: Carefully read the detailed data description.)

Solution

```
plot(x = Earnings_and_Height$height,
     y = Earnings_and_Height$earnings,
     pch = 16, # filled circle
     col = "black",
     xlim = c(45, 90),
     ylim = c(0, 85000),
     xlab = "height",
     ylab = "earnings",
     main = "E4.2 (c)")
```



The data documentation reports that individual earnings were reported in 23 brackets, and a single average value is reported for earnings in the same bracket. Thus, the dataset contains 23 distinct values of earnings.

d. Run a regression of *Earnings* on *Height*.
   i. What is the estimated slope?
   ii. Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

Solution

```
# regression
E42d <- lm(formula = earnings ~ height, data = Earnings_and_Height)

# i. estimated intercept, estimated slope
summary(E42d)
```

```
##
```

```
## Call:
## lm(formula = earnings ~ height, data = Earnings_and_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
```

```r
# predict value
E42d_predict <- function(x){
  E42d$coefficients %*% matrix(c(1, x), ncol = 1)

}

# ii. predict value: height = 67
E42d_predict(67)
```

```
##          [,1]
## [1,] 46901.26
```

```r
# ii. predict value: height = 70
E42d_predict(70)
```

```
##          [,1]
## [1,] 49024.28
```

```r
# ii. predict value: height = 65
E42d_predict(65)
```

```
##          [,1]
## [1,] 45485.92
```

e. Suppose height were measured in centimeters instead of inches. Answer the following questions about the *Earnings* on *Height* (in cm) regression.
   i. What is the estimated slope of the regression?
   ii. What is the estimated intercept?
   iii. What is the R^2?
   iv. What is the standard error of the regression?

Solution

```r
# translates from inches to cm
height_cm <- cm(Earnings_and_Height$height)

Earnings_and_Height <- cbind(Earnings_and_Height, height_cm)

# regression
E42e <- lm(formula = earnings ~ height_cm, data = Earnings_and_Height)
```

```r
# i. estimated slope # ii. estimated intercept
# iii. R^2 # iv. SE
summary(E42e)
```

```
##
## Call:
## lm(formula = earnings ~ height_cm, data = Earnings_and_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_cm     278.61      19.88  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
```

f. Run a regression of *Earnings* on *Height*, using data for female workers only.

  i. What is the estimated slope?

  ii. A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

Solution

```r
# female
Earnings_and_Height_f <- Earnings_and_Height[Earnings_and_Height$sex == 0, ]

# regression
E42f <- lm(formula = earnings ~ height, data = Earnings_and_Height_f)

# i. estimated slope # ii.
summary(E42f)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = Earnings_and_Height_f)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42748 -22006  -7466  36641  46865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12650.9     6383.7   1.982   0.0475 *
## height         511.2       98.9   5.169  2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07
```

A women who is one inch taller than average is predicted to have earnings that are $511.2 per year higher than average.

g. Repeat (f) for male workers.

Solution

```r
# male
Earnings_and_Height_g <- Earnings_and_Height[Earnings_and_Height$sex == 1, ]

# regression
E42g <- lm(formula = earnings ~ height, data = Earnings_and_Height_g)

# i. estimated slope # ii.
summary(E42g)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = Earnings_and_Height_g)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50158 -22373  -8118  33091  59228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43130.3     7068.5  -6.102  1.1e-09 ***
## height        1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16
```

A man who is one inch taller than average is predicted to have earnings that are $1306.9 per year higher than average.

h. Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, ui has a conditional mean of 0 given $Height$ (X_i)? (You will investigate this more in the Earnings and Height exercises in later chapters.)

Solution

Height may be correlated with other factors that cause earnings. For example, height may be correlated with "strength," and in some occupations, stronger workers may by more productive. There are many other potential factors that may be correlated with height and cause earnings and we will investigate of these in future exercises.