Empirical Exercise - E5.3

Chi-Yuan Fang

2021-03-21

On the text website, http://www.pearsonglobaleditions.com, you will find the data file **Birthweight_Smoking**, which contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. A detailed description is given in **Birthweight_Smoking_Description**, also available on the website. In this exercise, you will investigate the relationship between birth weight and smoking during pregnancy.

a. In the sample:
   i. What is the average value of *Birthweight* for all mothers?
   ii. For mothers who smoke?
   iii. For mothers who do not smoke?

Solution

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

E53a <- function(x){
  # i. sample mean
  mu <- mean(x)

  # ii. standard error = sample standard deviation (s) / sqrt(n)
  se <- sd(x)/sqrt(length(x))

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

  # iii. 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
```

```
E53a(Birthweight_Smoking$birthweight)
```

```
##       Mean Standard Error 95% Confidence Interval
## 1 3382.934       10.81137   3361.7352 - 3404.1321
```

```
# ii. # iii.
tapply(Birthweight_Smoking$birthweight, Birthweight_Smoking$smoker, E53a)
```

```
## $`0`
##      Mean Standard Error 95% Confidence Interval
## 1 3432.06       11.88903   3408.7462 - 3455.3737
##
## $`1`
##       Mean Standard Error 95% Confidence Interval
## 1 3178.832       24.04206   3131.6117 - 3226.0515
```

b. i. Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.
ii. What is the standard error for the estimated difference in (i)?
iii. Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

Solution

```
# data: non-smoker
Birthweight_Smoking0 <- Birthweight_Smoking[Birthweight_Smoking$smoker == 0,]

# data: smoker
Birthweight_Smoking1 <- Birthweight_Smoking[Birthweight_Smoking$smoker == 1,]

E53b <- function(x1, x2){
  # mean
  mu1 <- mean(x1); mu2 <- mean(x2)

  # s
  SD1 <- sd(x1); SD2 <- sd(x2)

  # n
  n1 <- length(x1); n2 <- length(x2)

  # i. # difference in mean
  mu <- mu2 - mu1

  # ii. # difference in standard error
  se <- sqrt(SD1^2/n1 + SD2^2/n2)

  # iii. # 95% confidence interval
  lower <- round(mu - qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  upper <- round(mu + qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  CI <- paste(lower, "-"  ,upper)

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

  Table
```

```
}

E53b(Birthweight_Smoking0$birthweight, Birthweight_Smoking1$birthweight)
```

```
##        Mean Standard Error 95% Confidence Interval
## 1 -253.2284       26.82106     -305.7967 - -200.66
```

c. Run a regression of *Birthweight* on the binary variable *Smoker*.
   i. Explain how the estimated slope and intercept are related to your answers in parts (a) and (b).
   ii. Explain how the $SE(\beta_1)$ is related to your answer in b(ii).
   iii. Construct a 95% confidence interval for the effect of smoking on birth weight.

Solution

```
E53c <- lm(birthweight ~ smoker, data = Birthweight_Smoking)

summary(E53c)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker, data = Birthweight_Smoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3432.06      11.87 289.115   <2e-16 ***
## smoker       -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

i. The intercept is the average birthweight for non-smokers ($Smoker = 0$). The slope is the difference between average birthweights for smokers ($Smoker = 1$) and non-smokers ($Smoker = 0$).

ii. They are the same.

iii. This is same as in b(iii).

d. Do you think smoking is uncorrelated with other factors that cause low birth weight? That is, do you think that the regression error term—say, u_i—has a conditional mean of 0 given *Smoking* (X_i)? (You will investigate this further in *Birthweight* and *Smoking* exercises in later chapters.)

Solution

Yes.