APSTA-GE 2003: Intermediate Quantitative Methods

Sample Solution - Assignment 1

Created on: 11/09/2020

Modified on: 11/11/2020

```
# Dependencies ----------------------------------------------------------------
library(car)
```

Instructions

In this part, you will conduct a simple linear regression analysis using the `lung capacity` data set. The dataset is available in `.csv` format, you will need to import the data into R using the function: `read.csv()`.

The dataset, `lung_capacity0.csv`, is in the data folder. Remember to setup correct working directory before importing the data set into R.

```
dat <- read.csv("../data/lung_capacity0.csv")

# Check the structure of `dat`
str(dat)
```

```
## 'data.frame':    80 obs. of  6 variables:
##  $ Sex           : int  1 1 1 0 1 1 1 1 1 0 ...
##  $ Height        : num  66.9 67.1 70.7 70.9 73.3 76.1 70.9 71.9 71.1 66.7 ...
##  $ Smoker        : int  0 0 0 0 0 0 0 1 1 0 ...
##  $ Exercise      : int  22 23 21 18 31 39 26 0 18 22 ...
##  $ Age           : int  49 41 82 26 69 42 68 40 35 50 ...
##  $ LungCapacitycc: int  5093 5116 5550 5530 5929 6212 5723 5133 5415 5075 ...
```

First, load the dataset to R using `read.csv()` as shown above. Then, check dimensions and the structure:

`dat` has 80 rows and 6 columns:

- `Sex`: Biological sex, binary, 1 as *male*, 0 as *female*
- `Height`: Height, numeric
- `Smoker`: Smoking status, binary, 1 as *smoker*, 0 as *non-smoker*
- `Exercise`: Number of hours for exercise per week, numeric
- `Age`: Age, numeric
- `LungCapacitycc`: Lung capacity in cc, numeric

Question 1

Report the sample size for this dataset.

Answer: Q1

```r
nrow(dat)
```

```
## [1] 80
```

There are 80 sample points.

Question 2

For each variable in the dataset, indicate whether it is numerical/quantitative or categorical.

Tip: please look at the values each variable takes.

A. Categorical

B. Numerical/Quantitative

Answer: Q2

Sex: A. Categorical

Height: B. Numerical/Quantitative

Smoker: A. Categorical

Exercise: B. Numerical/Quantitative

Age: B. Numerical/Quantitative

LungCapacitycc: B. Numerical/Quantitative

Question 3

Fill out the following summary statistics for lung capacity of smokers and non-smokers (round to the nearest whole number).

Note: for this part, you may assume that 0 corresponds to non-smoker and 1 corresponds to smoker.

	N	mean	variance	SE
smokers	A	B	C	D
nonsmokers	E	F	G	H

Answer: Q3

```r
# Create a function to perform rounding to the nearest whole number
rd <- function (n) {
  round(n, digits = 0)
}
# Alternative, you can use the round( ) function directly
# round(3.14, digits = 0)

# Smoker = 1; non-smoker = 0
# N
n_lung_smo <- nrow(dat[dat$Smoker == 1, ])
n_lung_non <- nrow(dat[dat$Smoker == 0, ])

# mean
avg_lung_smo <- mean(dat$LungCapacitycc[dat$Smoker == 1])
avg_lung_non <- mean(dat$LungCapacitycc[dat$Smoker == 0])

# Variance
var_lung_smo <- var(dat$LungCapacitycc[dat$Smoker == 1])
var_lung_non <- var(dat$LungCapacitycc[dat$Smoker == 0])

# SE
se_lung_smo <- sqrt(var_lung_smo / n_lung_smo)
se_lung_non <- sqrt(var_lung_non / n_lung_non)

cat(paste("",  "A: ", rd(n_lung_smo),   "\n",
               "B: ", rd(avg_lung_smo), "\n",
               "C: ", rd(var_lung_smo), "\n",
               "D: ", rd(se_lung_smo),  "\n",
               "E: ", rd(n_lung_non),   "\n",
               "F: ", rd(avg_lung_non), "\n",
               "G: ", rd(var_lung_non), "\n",
               "H: ", rd(se_lung_non)
    )
)
```

```
##  A:  36
##  B:  4951
##  C:  97558
##  D:  52
##  E:  44
##  F:  5552
##  G:  86395
##  H:  44
```

Question 4

Do smokers and non-smokers have the same lung capacity?

Translate this research question into null and alternative hypotheses that can be tested using a two sample T-test.

Answer: Q4

Research question: do smokers and non-smokers have equal lung capacity?

Null hypothesis (H_0): mean lung capacity for smokers is the **same** as non-smokers.

Alternative hypothesis (H_1): mean lung capacity for smokers is **different** from non-smokers.

Question 5

Conduct a Levene's test to determine whether a two-sample T-test with equal variance or unequal variance should be used.

Answer: Q5

```
leveneTest(LungCapacitycc ~ factor(Smoker), data = dat)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.4802 0.4904
##       78
```

Here, we need to first convert the dummy variable `Smoker` to factor using `factor(Smoker)`. Then, we pass it to the `leveneTest( )` to test equal variance.

Null hypothesis (H_0): two groups have equal variance.

Alternative hypothesis (H_1): two groups have different variances.

The F-value measures the distance of our sample from the null hypothesis (F-value = 0). The p-value, `Pr(>F)`, measures the probability of accepting the null. Since the p-value is larger than 0.05, we cannot reject the null. This indicates that two groups have equal variance.

Question 6

For this question, round all answers to the nearest tenth (i.e., one decimal place).

Run a two sample T-test comparing lung capacity of smokers vs. non-smokers.

The sample difference in mean lung capacity between smokers and non-smokers (non-smokers - smokers) is: ____

According to the two sample t test, the test statistic (t score) of the test is: ____

The p-value of this test is: ____

Answer: Q6

```
# Update our round function to perform rounding to the nearest tenth
rd <- function (n) {
  round(as.numeric(n), digits = 1)
}
# Alternative, you can use the round( ) function directly
# round(3.14, digits = 1)
```

```
rd(avg_lung_non - avg_lung_smo)
```

```
## [1] 601.3
```

The sample difference in mean lung capacity between smokers and non-smokers (non-smokers - smokers) is 601.3.

```
t.test(LungCapacitycc ~ Smoker, data = dat, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  LungCapacitycc by Smoker
## t = 8.8498, df = 78, p-value = 2.112e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  466.0219 736.5539
## sample estimates:
## mean in group 0 mean in group 1
##        5552.455        4951.167
```

According to the two sample t test, the test statistic (t score) of the test is 8.8.

The p-value is 0.

Question 7

There is a statistically significant association between smoking status and lung capacity.

Answer: Q7

This is true because the p-value of our T-test is smaller than 0.05. With more than 95% of confidence, we reject the null hypothesis. There is a statistically significant association between smoking status and lung capacity.

Question 8

Run a regression using lung capacity as the dependent variable (DV) and smoker as the independent variable (IV). Answer the following true or false question.

The intercept of the regression corresponds to mean lung capacity among non-smokers, and slope corresponds to the difference in mean lung capacity between smokers and non-smokers.

Answer: Q8

```r
mod_Q8 <- lm(LungCapacitycc ~ Smoker, data = dat)
summary(mod_Q8)
```

```
##
## Call:
## lm(formula = LungCapacitycc ~ Smoker, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -717.17 -262.17   -1.95  208.55  708.55
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5552.45      45.58  121.82  < 2e-16 ***
## Smoker       -601.29      67.94   -8.85 2.11e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302.3 on 78 degrees of freedom
## Multiple R-squared:  0.501,  Adjusted R-squared:  0.4946
## F-statistic: 78.32 on 1 and 78 DF,  p-value: 2.112e-13
```

This is true.

Question 9

Run a second regression using lung capacity as the DV and height as the IV.

Note: height is given in inches in this dataset. Round to the nearest tenth.

Report the intercept: _____

Report the slope: _____

Answer: Q9

```
mod_Q9 <- lm(LungCapacitycc ~ Height, data = dat)
summary(mod_Q9)
```

```
##
## Call:
## lm(formula = LungCapacitycc ~ Height, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1101.74  -118.81    17.92   146.02   965.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -757.271    615.950  -1.229    0.223
## Height        88.798      9.045   9.818 2.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286.2 on 78 degrees of freedom
## Multiple R-squared:  0.5527, Adjusted R-squared:  0.547
## F-statistic: 96.39 on 1 and 78 DF,  p-value: 2.827e-15
```

```
intercept_mod_Q9 <- coefficients(mod_Q9)["(Intercept)"]
slope_mod_Q9 <- coefficients(mod_Q9)["Height"]
```

The intercept: -757.3.

The slope: 88.8.

Question 10

Write a one sentence interpretation of the slope coefficient that you reported in the previous question.

Answer: Q10

For two groups of people whose average height differs by one inch, we expect the taller group to have lung capacity that is 88.8 cc higher on average.

Question 11

Create a new variable, `height0`, by centering height:

`height0 = height - sample mean height`

Note: variable "`height0`" records each subject's height difference from the mean height.

Run a third regression using lung capacity as the DV and `height0` as the IV. Answer the following true or false question.

The intercept of this regression corresponds to the mean lung capacity for those whose height is at the sample mean height.

Answer: Q11

```
dat$height0 <- dat$Height - mean(dat$Height)
mod_Q11 <- lm(LungCapacitycc ~ height0, data = dat)
summary(mod_Q11)
```

```
##
## Call:
## lm(formula = LungCapacitycc ~ height0, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1101.74  -118.81    17.92   146.02   965.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5281.875     32.002 165.046  < 2e-16 ***
## height0       88.798      9.045   9.818 2.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286.2 on 78 degrees of freedom
## Multiple R-squared:  0.5527, Adjusted R-squared:  0.547
## F-statistic: 96.39 on 1 and 78 DF,  p-value: 2.827e-15
```

This is true.

Question 12

The slope coefficients of the second and third regressions are the same because centering the IV does not change the effect of IV on the DV. It only affects the intercept of the regression line.

```
coefficients(mod_Q9)
```

```
## (Intercept)      Height
##  -757.27084    88.79791
```

```
coefficients(mod_Q11)
```

```
## (Intercept)     height0
##  5281.87500    88.79791
```

This is true.

Question 13

Create a new variable, `height1`, where

`height1 = height0 / SD(height0)`

Note: variable “`height1`” now records the height difference from mean height in terms of standard deviation of height.

Then, re-run the regression with lung capacity as the DV and `height1` as the IV. Report the intercept and slope of this regression (round to the nearest tenth):

Intercept: ____

Slope: ____

Answer: Q13

```
dat$height1 <- dat$height0 / sd(dat$height0)
mod_Q13 <- lm(LungCapacitycc ~ height1, data = dat)
summary(mod_Q13)
```

```
##
## Call:
## lm(formula = LungCapacitycc ~ height1, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1101.74  -118.81    17.92   146.02   965.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5281.9       32.0 165.046  < 2e-16 ***
## height1        316.2       32.2   9.818 2.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286.2 on 78 degrees of freedom
## Multiple R-squared:  0.5527, Adjusted R-squared:  0.547
## F-statistic: 96.39 on 1 and 78 DF,  p-value: 2.827e-15
```

```
intercept_mod_Q13 <- coefficients(mod_Q13)["(Intercept)"]
slope_mod_Q13 <- coefficients(mod_Q13)["height1"]
```

The intercept is 5281.9.

The slope is 316.2.

Question 14

For the regression above (using `height1` as the IV and lung capacity as the DV), the slope coefficient represents the expected difference in average lung capacity for two groups of people whose average height differs by ____.

Answer: Q14

Since `height1` is the standardized `Height`, the measurement unit of the slope coefficient will then be **one standard deviation**.

Question 15

For the regression above (using `height1`), the intercept of this regression is the same as that from the regression using `height0` because it represents the expected mean lung capacity of people with average height in this dataset.

Answer: Q15

```
coefficients(mod_Q11)
```

```
## (Intercept)     height0
##  5281.87500    88.79791
```

```
coefficients(mod_Q13)
```

```
## (Intercept)     height1
##    5281.875     316.177
```

This is true.

Question 16

Create a new variable, `height2`, that measures subjects centered height (`height0`) in the unit of "centimeter".

Variable "`height2`" now measures how much each subject's height is different from the mean height in the unit of centimeter.

Then, re-run the regression with lung capacity as the DV and `height2` as the IV.

Hint: 1 inch = 2.54 centimeter, hence `height2 = height0 * 2.54`.

Report the intercept and the slope of this regression (round to the nearest tenth):

Intercept: ____ Slope: ____

Answer: Q16

```
dat$height2 <- dat$height0 * 2.54
mod_Q16 <- lm(LungCapacitycc ~ height2, data = dat)
summary(mod_Q16)
```

```
##
## Call:
## lm(formula = LungCapacitycc ~ height2, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1101.74  -118.81    17.92   146.02   965.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5281.875     32.002 165.046  < 2e-16 ***
## height2       34.960      3.561   9.818 2.83e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 286.2 on 78 degrees of freedom
## Multiple R-squared:  0.5527, Adjusted R-squared:  0.547
## F-statistic: 96.39 on 1 and 78 DF,  p-value: 2.827e-15
```

```
intercept_mod_Q16 <- coefficients(mod_Q16)["(Intercept)"]
slope_mod_Q16 <- coefficients(mod_Q16)["height2"]
```

The intercept is 5281.9.

The slope is 35.

Question 17

For the regression above (using `height2` as the IV and lung capacity as the DV), the intercept of this regression is the same as that from the regression using `height0` because it represents the expected mean lung capacity of people with average height in this dataset.

Answer: Q17

```r
coefficients(mod_Q11)
```

```
## (Intercept)     height0
##  5281.87500    88.79791
```

```r
coefficients(mod_Q16)
```

```
## (Intercept)     height2
##  5281.87500    34.95981
```

This is true.

Question 18

For the regression above (using `height2`), the slope coefficient represents the expected difference in average lung capacity for two groups of people whose average height differs by ________.

Answer: Q18

Refer to Question 16. The measurement unit of `height2` is **centimeter**.

END: Sample Solution - Assignment 1