Homework 2

BANA7038

Write a report to answer the questions in this homework. When answering each question, please organize your answer to have the following 4 parts:

1. What you are trying to do? (Your goal)

2. R code (How do you realize it?).

3. R output (What is the output from R?).

4. Your observations (What do you observe from the output? How do you interpret the output?)

If necessary, you can repeat 1 through 4 for many times to answer one question fully.

**Instructions on the report:**

1. Avoid printing large tables. Avoid printing the entire data set or confidence intervals or too many number at once, instead, visualize them in figures and show only a few rows.

2. Avoid plotting large figures. Use par() to plot multiple figures in one panel to save space.

Plot square figure, do not generate “skinny and tall”, and “short and wide” figures.

3. Limit the length of your report. Try to be as concise as possible.

4. Mark the question numbers in your report in **LARGE and BOLD** font.

5. Separately write down your last name, first name (no abbreviations), and M number. For example:

Last Name: Smith

First Name: John

M-number: M12345678

1. Read <tombstone.csv> into R. Use response variable = Marble Tombstone Mean Surface Recession Rate, and covariate = Mean SO2 concentrations over a 100 year period. Description: Marble Tombstone Mean Surface Recession Rates and Mean SO2 concentrations over a 100 year period.

2. Plot data and briefly describe what you observe, i.e., positive trend or negative trend and others.

3. Perform linear regression using lm() function

3.1. Obtain coefficient estimates , . What do these estimate imply in terms of the relationship between the response and the covariate?

3.2. Obtain fitted values and the sum of fitted values. What do these fitted values represent?

3.3. Obtain the sum of all values of response variable.

3.4. Verify the fact that the sum of fitted values is always the same as the sum of response variable. In addition, verify the fact that the mean of the fitted values is always the same as the mean of response variable, .

3.5. Obtain residuals and the sum of residuals, and verify the fact that the sum of residuals is always zero.

3.6. Obtain the standard errors of , . Are these standard errors satisfactory and why?

4. Suppose we increase SO2 Concentration by one unit, how does such a change influence the Surface Recession Rate?

5. Does the intercept of the linear regression have natural interpretation? If so, what does it mean?

6. Which city (i.e., observation) has the highest Surface Recession Rate?

7. Which city (i.e., observation) has the largest residual (i.e., the largest absolute value) according to the linear regression you just fitted?

8. Calculate the mean of covariate and mean of response. Verify the fact that the fitted regression line go through the point .

9. Repeat the same questions (1 to 8, except 6, 7) for the date set <bus.csv>. Description: Cross-sectional analysis of 24 British bus companies (1951). Use response variable = Expenses per car mile (pence), covariate = Car miles per year (1000s).