Homework 3

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. Obtain , explain what it means.

3. Perform the following hypothesis testing and interval estimation using lm() and other related R functions.

3.1. Perform hypothesis tests for each coefficient, obtain p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

3.2. Compute interval estimation for coefficients, interpret the meanings of these quantities.

3.3. Make prediction of the Recession Rate at the 1st, 2nd, 3rd, …., 99th percentiles of the SO2, and obtain their prediction intervals (two types of the intervals).

3.4. Plot data points, the regression line, the predictions and its intervals (two types) and show that the interval is wider on both sides and narrow in the center.

3.5. Making prediction at what range of values of SO2 would be considered extrapolation? Is it OK to do extrapolation in this case?

4. Repeat the same questions (1-3) 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).