HW 7

Consider data collected in the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The SUPPORT dataset contains information on a random sample of 1000 patients admitted into one of five academic medical centers across the US. While the original goal of the SUPPORT study was to identify factors that correlate with poor outcomes following hospital admission, it has also been used to examine factors that affect hospital costs. Our goal is to estimate the association between patient age and total hospital costs using the SUPPORT data.

The variables in the dataset include age (age), sex (sex), length of hospital stay (slos), disease causing hospitalization (dzgroup), number of comorbidities (num.co), education level of the patient (edu), income of the patient (income), charges by the hospital (charges), Medicare adjusted charges (totcst), and race (race) of the patient.

Focus on style, clarity, and readability of your document (i.e., check your pdf/doc before submitting).

Part 1: Data Cleaning

**Before beginning this homework, install the package Hmisc.*

1. In the setup chunk:
    a. Add `message = FALSE` and `warning = FALSE` to the chunk options and change `echo = FALSE` to `echo = TRUE`;
    b. Load all required packages and the dataset in this chunk *(Tip: If reading in the data using the readr package, use the function read_delim(). This will eliminate any problems with quotation marks.)*
2. Using any of the methods we have learned in class, clean the dataset by:
    a. Renaming the columns `slos`, `dzgroup`, `num.co` and `totcst` to `length_stay`, `disease_group`, `num_comorbid`, and `total_cost`, respectively;
    b. Changing all categorical variables to factors (and capitalizing each level using the `capitalize()` function from `Hmisc`); and
    c. Creating a new variable called `log_total_cost` that is the (natural) log of `total_cost`.

HW 7

Part 2: Exploratory Data Analysis

3. Use the following code to look for patterns of missingness in the data.

```
Unset
na_patterns <- naclus(support)
naplot(na_patterns, "na per var")
plot(na_patterns)
```

4. Comment on both plots in a few complete sentences. Focus especially on possible relationships between variables with missing values. Why does it make sense that these variables would be missing? *This question requires you to think beyond the plots and statistics and focus on the participants and the types of questions the researchers asked them.*
5. Plot 1: Create a histogram of the distribution of `total_cost`. You may use any plotting functions from any R package. Be sure to include proper x and y-axis labels and a title.
6. Comment on the shape, center, and spread of Plot 1.
7. Plot 2: Construct a scatterplot of `total_cost` vs. `age`. You may use any plotting functions from any R package. Be sure to include proper x and y-axis labels and a title.
8. What is a potential confounding variable (in the dataset) in the relationship between age and total cost? Explain how it is *both* related to the response and the explanatory variable.

Bonus: plot the relationship between total cost, age, and your confounder.

Part 3: Data Analysis

9. Write the hypothesis to see if there is a linear relationship between total cost and age in both symbols and words.
10. Check the necessary model assumptions for performing a simple linear regression analysis. Explain which conditions are not satisfied and how you know.
11. Instead, model the relationship between the log of total cost and age. Check these assumptions. *Note you do not have to change your hypotheses.*
12. Perform the remaining three steps of the hypothesis test, remembering not to show any R code, only necessary output. *Suggestion: inline R code is a great way to print results!*
13. Interpret the slope coefficient in the context of the problem, remembering that the responses are on the log scale.

Part 4: Discuss the Results

14. Is there evidence that age is linearly associated with (log) total cost? If yes, describe the relationship in complete sentences using your previous results. If no, state the evidence you used from your previous results to come to this conclusion.