Homework 5 Graded

Sheridan Grant

Must be uploaded to Canvas under "Homework 5 Graded" by
Monday, May 4 at 11:59pm

Instructions

Format your .RMD file using the template on the course website. **Submit the .RMD file, the knitted .html output, and any other files or folders needed as a single .zip file.**

The grader will be compiling your .RMD file and making sure it knits. Any libraries/packages needed should be near the top of the .RMD file, so the grader can make sure they're installed. Any other files needed to knit the .html should be in the zipped folder you turn in. **If your code does not knit and there is no immediate fix, the grader will grade your HTML for a [-10pts] penalty.**

Any time I ask you to demonstrate something, show something, generate something, etc., you must provide the code and/or text commentary that does so.

Finally, we will be giving [5pts] for code style and cleanliness. For any function you write, include a comment on the line above the function saying what the function expects as input and what it outputs. If you do this and the rest of your code is reasonably neat then this is an easy [5pts].

1 Programming Puzzles

For this homework, you'll need the SAT and COVID-19 data that are on the course website. I refer to the SATSum variable as the "combined score."

(a) Write a function named `greet` that takes in a single argument `greeting`, and if `greeting` is an upper-case character string, returns "Hello, [`greeting`]." I.e. if `greeting == 'SHERIDAN'`, it returns `''Hello, SHERIDAN''`. However, if `greeting` is not an upper-case character string, it should "throw an error"—not just print a message, but an actual error. There are easy built-in functions that will help. You may assume `greeting` is upper-case if `toupper` doesn't change `greeting` in any way. [3pts]

(b) Redo question 1(b) from HW 4 Graded, but don't use `t.test` or your own t test function. Instead, use a categorical variable as the covariate in two `lm` models. Show the model summaries and print out the two p-values separately. You may use `t.test` to verify that the p-values are correct. [3pts]

(c) Write code that generates data from the following model: $X \sim \mathrm{N}(0,1)$, $Z \sim \mathrm{N}(X,1)$, $Y \sim X + 2Z - XZ + \epsilon$ where $\epsilon \sim \mathrm{N}(0,1)$ and generate 10^5 data points from this model. Fit two models to this data: model 1, $Y_i = \beta_0 + \beta_X X_i + \beta_Z Z_i + \epsilon_i$; and model 2, $Y_i = \beta_0 + \beta_X X_i + \beta_Z Z_i + \beta_{XZ} XZ\epsilon_i$. For each model, find 95% confidence intervals for β_X and β_Z, and identify whether or not they contain the true values of β_X and β_Z. Write a sentence summarizing your results in terms of "model misspecification." [4pts]

2 Inference in Linear Models

In this problem, you'll do inference on linear model coefficients and predictions like we did towards the end of class Wednesday. The questions are designed to be similar to code I wrote for lecture10.RMD, posted on the website.

(a) Fit the model `FYGPA ~ SATSum + HSGPA`. Provide a 90% confidence interval for the first-year college GPA of an "average" student with SAT 100 and HS GPA 3.0. [2pts]

(b) Now fit the model `FYGPA ~ HSGPA*sex` (with `sex` encoded as `male == 0` and `female == 1`). Make a scatterplot of HS vs. college GPA with male students colored red and female students colored blue. Add two lines to the scatterplot: the line of predicted outcomes for female students, again in blue, and the line of predicted outcomes for male students, again in red. [4pts]

(c) Find a 90% confidence interval for the change in first-year college GPA associated with a 1-point increase in HS GPA *for female students*. [3pts]

(d) Write a function `infer` that takes as inputs:

   (a) A linear model object `lmod`,

   (b) A numeric vector `a`, and

   (c) A confidence level `conf` between 0 and 1

   and returns a `100conf%` interval for $a^T\beta$, where $\beta = [\beta_0, \beta_1, \ldots, \beta_d]$ is the vector of model coefficients. You may assume that a and β have the same length.[1] You may **not** use the `confint` function to write `infer`, but you can use whatever

[1] If you've done Q1(a) it'll be easy for you to check for this and throw an error if it's not true, but we won't be grading you based on this.

you like to check your work. The lecture 10 code, the posted lecture slides from April 29, and checking your work with `confint` should help you. [6pts]

3 Linear Models for COVID-19

(a) Using the COVID-19 data (on the website), create a data frame with two columns, `Cases` and `Date`, where `Cases` is confirmed cases and `Date` is a proper date-type variable, and where the rows correspond to all days in March with a non-zero number of confirmed cases in the city of San Francisco. There should be 27 rows. Print the first 6 rows of this data frame with `head` for the grader. [3pts]

(b) You want to model how the number of cases in SF evolved over the month of March. Plot the number of confirmed cases over time. Why is a linear model inappropriate ("inappropes," as the kids say these days) here? Find a transformation of the `Cases` variable for which a linear model appears more appropriate, and plot *it* over time. [3pts]

(c) Fit a linear model to the transformed outcome and the Date variable. [1pt] Interpret the coefficient of interest. [1pt] Show the residual plot and comment on how plausible the linear model assumptions are, and show the Normal Q-Q plot and comment on how trustworthy inferences about model coefficients will be. [1pt] What is the model's prediction for the number of cases on April 1 in SF? How does this compare to the number there actually were? [1pt]