Stat 021 Homework 4

Suzanne Thornton

Due: Friday, Oct. 25, 12:00pm

Instructions: A **pdf** version of your homework must be submitted to Gradescope by **noon** on the due date. The course passcode is **MPKJ4Z**. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will receive a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q 1) Recall the skyscraper data set used in Homework 3 ("skyscraper_data.txt"). This data investigates how the height (in meters) of a skyscraper depends on the number of floors it has. (5 points)

a) Suppose a developer is working on a new building that has taken the 15 years to get the go-ahead. Suppose they are cheekily designing the building to have 15 floors, one for each year of struggle to get the building approved. If the architect needs to know how tall this building may be, would you use a prediction interval or a confidence interval? Justify your answer.

b) As we did in class, create a scatter plot of the observed data and overlay on this plot the estimated regression line and the confidence and prediction bands.

Q 2) Again, referencing the skyscraper data in Q1, note that there is a categorical variable called "purpose" included in the data set. Suppose we are interested in determining if there is a significant difference in the average height of a building depending on what it's purpose is. Using height (in meters) as the response and purpose as the explanatory variable, fit an ANOVA model to this data after excluding the data point for the only hospital. What does the result of the ANOVA F-test indicate? (5 points)

Q 3) We all know that the significance level (α) represents the probability of a false positive (i.e. a type I error) in our inference problem. Related to this concept is the probability of correctly detecting a positive. In statistics, this probability is called the power of the study and is often denoted as $1 - \beta$ where β is the probability of a type II error. (Hence the power is the probability of NOT making a type II error.) What if we wanted to collect more skyscraper data to increase the power of our test in Q 2? Using this example as a guide: https://stats.idre.ucla.edu/r/dae/one-way-anova-power-analysis/ and assuming we can collect enough data so that all categories for the variable "purpose" have the same number of observations (i.e. we have a balanced ANOVA design), how many more observations of hospitals, hotels, offices, and residential buildings specifically do we need to achieve 85% power? (5 points) Hint: for this problem, you can assume a balanced ANOVA design for the power analysis part but your recommendations for sample size don't have to result in a balanced design.

Q 4) Suppose we are interested in studying the effectiveness of the recycling/composting programs at Swarthmore. I.e. we are investigating the waste that is disposed in the trash/recycle/compost bins across campus. Use your imagination to come up with three different research questions related to this topic in the case where

1. We have two numerical variables of interest;
2. We have one numerical variable of interest and one categorical variable of interest;
3. We have two categorical variables of interest.

Please be sure to clearly state what are your variables, what roles they play, and the research question. Each research question you come up with should be answerable by one of: a simple linear regression, an ANOVA model, or a chi-squared test. (5 points)