Stat 021 Homework 3

Suzanne Thornton

Due: MONDAY, Sept 1, 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 recieve a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q 1) In a survey of 988 men aged 18˘24, the regression equation for predicting height from weight was: (3 points)

$$height = 62.4 + (0.047)(weight),$$

where height is measured in inches and weight is measured in lbs.

a) Is the following statement a correct interpretation of the regression line: "If someone gaines 10 pounds, he will get taller by (0.047)(10) = 0.47 inches"? If not, provide a better explaination for the meaning of the slope.

b) Suppose the variance of our model errors is $\sigma^2 = 2in$. What percentage of all 200-pound men are taller than 74 inches? (Assume the regression model assumptions are met.)

Q 2) Below are two Quantile-quantile plots of GRE General Test Verbal scores for students intending graduate study in psychology, classics, and economics. Here we are comparing the psychology test scores to the classics and to economics. We are interested in how the population psychology students compares to the population of classics students and to the population of economics student. How do you interpret the patterns (deviations from the diagonal lines) in these plots? (3 points)

Q 3) How does the height (Y) of a skyscraper depend on the number of stories (i.e. floors) it has (x)? This data (availabe on Moodle) was collected in 2018-2019 and recorded some information on different skyscrapers in NYC.(10 points)

a) Import the data from "skyscraper_data.txt" and make a scatterplot of height vs stories. Based on this scatterplot, does the relationship appear to be linear? Is the relationship a strong one?

b) Calculate the regression line that uses the number of stories a building has to predict the height of a building. What are: the equation of the regression line, the value of the standard deviation of height, and the value of R-squared?

c) Calculate the correlation between height and stories using the *cor()* function in base R. Interpret this number and describe the relationship between this number and the R-squared value in part (b).

d) Calculate a 95% confidence interval for the model parameter β_1, the slope of the variable "floors". How would you explain the meaning of this confidence interval, in the specific context of this dataset, to an architect who has never taken a statistics class?

e) Test the hypothesis that $\beta_1 = 0$ at an $\alpha = 0.05$ significance level. State your null and alternative hypotheses and report the test statistic and p-value (and/or rejection region). Interpret, in the context of the problem, the results of this test.

f) Plot the residuals from your linear model in part (b) (on the vertical axis) with the number of stories (on the horizantal axis). Are there any apparent violations of the regression model assumptions? Explain briefly.

g) Make a Normal probability plot to determine if the residuals look like they come from a Normal distribution. Interpret your results.

Q 4) Let's consider the data set called *msleep* which is contained in the R package ggplot2 and is similar to the *sleep* data set that we used in HW 1. This *msleep* data set looks at the amount of time spent sleeping for different mammals and records other factors such as brain and body weight. For this problem, we are interested in the total amount of sleep an animal gets (*sleep_total*) as predicted by the total body weight of the animal (*bodywt*).(4 points)

a) There is a qualitative variable named *order* in this data set. Fit two separate linear regression models for the animals of *order* "Carnivora" and of *order* "Primates".

b) What is the estimate of the variance of the errors for each of these two models?

c) Do you think We could fit a single linear regression model to both *orders* arnivora and Primates? Justify your answer with 1-2 sentences and (possibly) a supporting plot.