Homework 4

P8130 Fall 2022

Due: November 13, 2022 at midnight Eastern

P8130 Guidelines for Submitting Homework

- Your homework must be submitted through Courseworks. No email submissions!
- Only one PDF file should be submitted, including all derivations, graphs, output, and interpretations. When handwriting is allowed (this will be specified), scan the derivations and merge ALL PDF files (http: //www.pdfmerge.com/).
- You are encouraged to use R for calculations, but you must show all mathematical formulas and derivations. Please include the important parts of your R code in the PDF file but also submit your full, commented code as a separate R/RMD file.
- To best follow these guidelines, we suggest using Word (built in equation editor), R Markdown, Latex, or embedding a screenshot or scanned picture to compile your work.

DO NOT FORGET: You are encouraged to collaborate on homeworks, explain things to each other, and test each other's knowledge. But Do NOT hand out answers to someone who has not done any work. Everyone ought to have ideas about the possible answers or at least some thoughts about how to probe the problem further. Write your own solutions!

Problem 1 (10 points)

A new device has been developed which allows patients to evaluate their blood sugar levels. The most widely device currently on the market yields widely variable results. The new device is evaluated by 25 patients having nearly the same distribution of blood sugar levels yielding the following data:

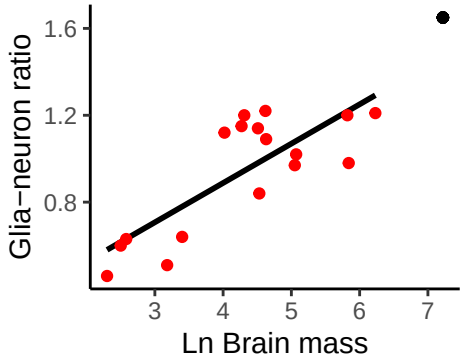
125 123 117 123 115 112 128 118 124 111 116 109 125 120 113 123 112 118 121 118 122 115 105 118 131

a) Is there significant ($\alpha = 0.05$) evidence that median blood sugar readings was less than 120 in the population from which the 25 patients were selected? Use the sign test and report the test statistic and p-value.

b) Is there significant ($\alpha = 0.05$) evidence that median blood sugar readings was less than 120 in the population from which the 25 patients were selected? Use the Wilcoxon signed-rank test and report the test statistic and p-value.

Problem 2 (15 points)

Human brains have a large frontal cortex with excessive metabolic demands compared with the brains of other primates. However, the human brain is also three or more times the size of the brains of other primates. Is it possible that the metabolic demands of the human frontal cortex are just an expected consequence of greater brain size? A data file containing the measurements of glia-neuron ratio (an indirect measure of the metabolic requirements of brain neurons) and the log-transformed brain mass in nonhuman primates was provided to you along with the following graph.



a) Fit a regression model for the nonhuman data using ln (brain mass) as a predictor. (Hint: Humans are "homo sapiens".)

b) Using the nonhuman primate relationship, what is the predicted glia-neuron ratio for humans, given their brain mass?

c) Determine the most plausible range of values for the prediction. Which is more relevant for your prediction of human glia-neuron ratio: an interval for the predicted mean glia-neuron ratio at the given brain mass, or an interval for the prediction of a single new observation?

d) Construct the 95% interval chosen in part (c). On the basis of your result, does the human brain have an excessive glia-neuron ratio for its mass compared with other primates?

e) Considering the position of the human data point relative to those data used to generate the regression line (see graph above), what additional caution is warranted?

Problem 3 (25 points)

For this problem, you will be using data `HeartDisease.csv`. The investigator is mainly interested if there is an association between 'total cost' (in dollars) of patients diagnosed with heart disease and the 'number of emergency room (ER) visits'. Further, the model will need to be adjusted for other factors, including 'age', 'gender', 'number of complications' that arose during treatment, and 'duration of treatment condition'.

a) Provide a short description of the data set: what is the main outcome, main predictor and other important covariates. Also, generate appropriate descriptive statistics for all variables of interest (continuous and categorical) – no test required.

b) Investigate the shape of the distribution for variable `totalcost` and try different transformations, if needed.

c) Create a new variable called `comp_bin` by dichotomizing 'complications': 0 if no complications, and 1 otherwise.

d) Based on your decision in part (b), fit a simple linear regression (SLR) between the original or transformed `totalcost` and predictor `ERvisits`. This includes a scatterplot and results of the regression, with appropriate comments on significance and interpretation of the slope.

e) Fit a multiple linear regression (MLR) with `comp_bin` and `ERvisits` as predictors.

   i) Test if `comp_bin` is an effect modifier of the relationship between `totalcost` and `ERvisits`. Comment.

   ii) Test if `comp_bin` is a confounder of the relationship between `totalcost` and `ERvisits`. Comment.

   iii) Decide if `comp_bin` should be included along with `ERvisits`. Why or why not?

f) Use your choice of model in part (e) and add additional covariates (age, gender, and duration of treatment).

   i) Fit a MLR, show the regression results and comment.

   ii) Compare the SLR and MLR models. Which model would you use to address the investigator's objective and why?