Homework 4

P8130 Fall 2025

Due: November 25, 2025 at 11:59pm

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.

REMINDER: You are encouraged to collaborate on homework, explain things to each other, and test each other's knowledge. But **everyone must complete their own assignment and write their 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 (10 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? For this problem, use the provided data file entitled "Brain data".

a) Using only the non-human data, make a scatterplot of (natural) log of brain mass on the X-axis and glia-neuron ratio as outcome. Then fit the corresponding regression model and write an expression for the fitted regression line.
b) Using the nonhuman primate relationship, what is the predicted glia-neuron ratio for humans, given their brain mass?
c) Construct a 95% prediction interval corresponding to the prediction made in part (c). Based on this, does the human brain have an excessive glia-neuron ratio for its mass compared with other primates?
d) Is there anything to be cautious about when using the non-human data to make predictions for humans? Explain your answer.

Problem 3 (20 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. The model may need to be adjusted for other factors, including age, gender, number of complications that arose during treatment, and duration of treatment condition.

a) Generate appropriate descriptive statistics for all variables of interest (continuous and categorical)

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

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

d) Fit a simple linear regression (SLR) model with `totalcost` (original or transformed, based on your decision in part (b)) as the outcome variable and `ERvisits` as predictor. This should include a scatterplot and results of the regression analysis, with appropriate comments on significance and interpretation of the slope estimate.

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

   I. Test for an interaction between `comp_bin` and `ERvisits`. Give your conclusions and interpret the results.

   II. Test whether `comp_bin` is a confounder of the relationship between `totalcost` and `ERvisits`. Give your conclusions and interpret the results.

   III. Should `comp_bin` be included as a covariate in the model (along with `ERvisits`)? Explain your reasoning.

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

   I. Fit a MLR, provide the fitted regression equation, and give the interpretation of each estimated parameter. Which variables seem to be significant?

   II. Compare the SLR and MLR models using the appropriate testing procedure. Which model would you use to address the investigator's objective and why?

.