P8130 Fall 2019: Biostatistical Methods I

Homework 4

Due, Nov 19 @ 5:00pm

P8130 Guidelines for Submitting Homework

Your homework should be submitted only through CourseWorks. No email submissions!

All derivations, graphs, output and interpretations to each section of the problem(s) must be included in the PDF (not the code), otherwise it will not be graded.

Only 1 PDF file should be submitted. When derivations were required and handwriting was allowed, scan the derivations and merge ALL PDF files (http://www.pdfmerge.com/) into a single one.

We are encouraged to use R for calculations, but you still have to show the mathematical formulae. Also, make sure to also submit your commented code as a separate R/RMD file.

DO NOT FORGET:

You are encouraged to collectively look for answers, explain things to each other, and use questions to test each other 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!

Problems 1 and 2 can be handwritten - legibly – scanned and incorporated into the HW PDF.

Problem 1 (10p)

Consider the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \ \varepsilon_i \sim N(0, \sigma^2).$$

The 'estimated errors' of the model are called *residuals* and denoted by $e_i = Y_i - \widehat{Y}_i$.

a) Write the *Least Squares* line equation and show that it always goes through the point $(\bar{X}, \bar{Y})$.
b) Show that $corr(e_i, \widehat{Y}_i) = 0$. What are some possible explanations of having a high correlation between residuals and fitted values?

Problem 2 (10p)

Consider the simple linear regression model: $\tilde{Y} = X\tilde{\beta} + \tilde{\varepsilon}, \tilde{\varepsilon} \sim N(0, \sigma^2 I)$.

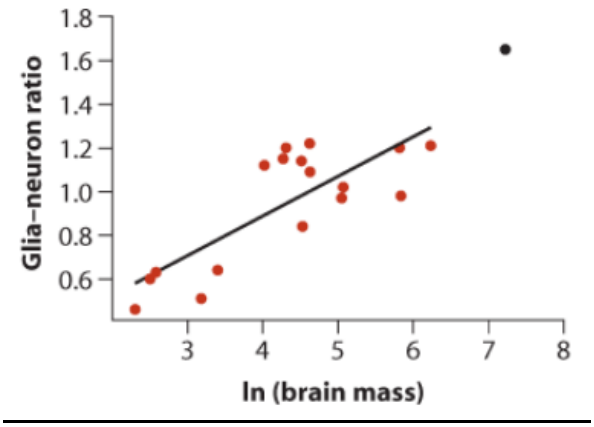
Using matrix notation:

a) Derive the Least Squares vector of estimates for the model coefficients and find its expected value.
b) Derive the estimated variance-covariance matrix (all four components) associated with model coefficients, i.e., $cov(\tilde{\beta})$.

For all problems below, assume a significance level of 0.05 unless stated otherwise. You can use R to perform the analyses, but you need to write the hypotheses where specified.
NO PICS will be allowed!

Problem 3 (15p)

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

Problem 4 (25p)

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 'total cost' 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 our decision in part b), fit a simple linear regression (SLR) between the original or transformed 'total cost' 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 ‘total cost’ and ‘ERvisits’. Comment.

ii) Test if ‘comp_bin’ is a confounder of the relationship between ‘total cost’ 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?