P8130 Fall 2018: Biostatistical Methods I

# Homework 4

Due, Nov 16th @ 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.

You are encouraged to use R for calculations, but you still have to show the mathematical formulae. For some problems, you can use R ONLY and the problem will specifically state that.

Also, make sure you include your commented code at the end of the document (PDF) or attach a separate R 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*

You are NOT supposed to 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 (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 (not derive) the *Least Squares* estimators of $\beta_0, \beta_1$ and show that they are unbiased estimators of the true model parameters – do not use matrix notation! (5p)
b) Write the *Least Squares* line equation and show that it always goes through the point $(\bar{X}, \bar{Y})$. (2p)
c) Use maximum likelihood method to derive an estimator of $\sigma^2$. Find its expected value and comment on the unbiasness property. (3p)

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.

Problem 2 (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. (5p)
b) Investigate the shape of the distribution for variable 'total cost' and try different transformations, if needed. (2p)
c) Create a new variable called 'comp_bin' by dichotomizing 'complications': 0 if no complications, and 1 otherwise. (1p)
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. (5p)
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. (2p)
    ii)Test if 'comp_bin' is a confounder of the relationship between 'total cost' and 'ERvisits'. Comment. (2p)
    iii) Decide if 'comp_bin' should be included along with 'ERvisits. Why or why not? (1p)

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. (5p)
    ii) Compare the SLR and MLR models. Which model would you use to address the investigator's objective and why? (2p)

Problem 3 (15p)

A hospital administrator wishes to test the relationship between 'patient's satisfaction' (Y) and 'age', 'severity of illness', and 'anxiety level' (data 'PatSatisfaction.xlsx'). The administrator randomly selected 46 patients, collected the data, and asked for your help with the analysis.

a) Create a correlation matrix and interpret your initial findings. (2p)
b) Fit a multiple regression model and test whether there is a regression relation. State the hypotheses, decision rule and conclusion. (3p)
c) Show the regression results for all estimated coefficients with 95% CIs. Interpret the coefficient and 95% CI associated with 'severity of illness'. (5p)
d) Obtain an interval estimate for a new patient's satisfaction when Age=35, Severity=42, Anxiety=2.1. Interpret the interval. (2p)
e) Test whether 'anxiety level' can be dropped from the regression model, given the other two covariates are retained. State the hypotheses, decision rule and conclusion. (3p)