STAT 614 - HW 7

Due: Monday, November 23, 2020 in Blackboard by 11:59pm.

Instructions: Please type your solutions in a separate document and upload the document in Blackboard as a pdf. I will not be collecting syntax for this assignment. You will need concepts from Chapters 7 & 8 on the simple linear regression model (in addition to past models!). HW 8 will address multiple regression.

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. The data set FEV.csv in Blackboard contains determinations of FEV for 654 children ages 3 through 19 who were seen in the Childhood Respiratory Disease (CRD) Study in East Boston, Massachusetts. These data are part of a longitudinal study to follow the change in pulmonary function over time in children. Variables in the data set are the participant ID number, Age (in years), FEV (in liters), Height (in inches), a binary Sex indicator (0 = female/1 = male), and Smoking status (0 = non-smoker/1 = current smoker).

1. Characterize the association between pulmonary function (FEV) and smoking status. To do this answer the following questions:

   a. Use the natural log transformation of FEV and examine an independent two-sample procedure to test for differences in the population mean LN(FEV) between the two smoking groups. Provide a brief summary of the model results. Is there evidence of an association between (transformed) pulmonary function and smoking status? If so, estimate the extent of the association (that is, give estimate and confidence interval for the parameter of interest).
   b. Are you surprised by the results in (a)? (Note that this is the *unadjusted* association.)

2. The smoking status variable is an *indicator* variable in that it takes the value 0 for non-smokers and 1 for current smokers. Non-smokers are considered the *reference* group. Even though this is a *categorical* (i.e. qualitative or grouping) variable, we can use indicator variables to designate groups in the regression procedure. (This is solely due to the 0/1 status of the variable!) Fit the simple linear regression model of FEV (use the natural log transformed FEV) with smoking status as the explanatory variable.

   a. Test the null hypothesis of no association between smoking status and FEV in the simple linear regression model.
   b. Interpret the slope coefficient.
   c. Compare your results to those in part 1a. You should draw *identical* conclusions. Do you? (Same estimated difference in mean LN(FEV) between non-smokers and smokers, same confidence interval, same p-value.)
   d. Explain how to use a regression model with indicator variables to include a three (or more) category explanatory variable. For example, *if* smoking status was 0 for never smoked, 1 for past smoker, and 2 for current smoker, how would you incorporate the three smoking levels into a regression model? (Note: This gives us a way to incorporate both quantitative and qualitative variables into a model!)