HW #3
BIOSTAT 705 Spring 2018
Due on 4/10

1. Suppose we collected measurements on 70 subjects, where the response (Y) was SBP, and the predictors were age (X as continuous), gender (Z as categorical 0=male, 1=female), education level (high-school, BS, MS and PhD).

   a) Using dummy variables, specify the appropriate single regression model relating SBP to age, gender, and education level includes the following interactions: age×gender, age×education, and gender×education.

   b) In terms of regression parameters from a), state the null hypothesis to be tested for the following:

      1) All regression equations are coincide.
      2) All regression equations are parallel.
      3) All gender equations are parallel, controlling for education level.
      4) All education equations are parallel, controlling for gender.

   c) For each of the hypotheses given in part a) specify the degrees of freedom and the test statistic.

2. An investigator is interested in whether Bayer aspirin (treatment "A"), Tylenol (acetaminophen) (treatment "B") or Aleve (naproxen) (treatment "C") works more quickly to relieve the pain of a common headache. She recruits n individuals with frequent headaches, randomly assigns them to one of the three pain killers, asks them to take the medication upon first signs of the headache, and to record the time until the pain is gone (Y).

   a) Using indicator (dummy) variables in regression and set up the model for the experiment above. Assume n_i $i = 1, 2, 3$ subjects are randomized to group i, where $n = n_1 + n_2 + n_3$.

   b) Write your model in part a) in a matrix form as $Y = X\beta + \epsilon$, identify the design-matrix X as well as Y, β and ϵ vectors.

   c) State the null hypothesis of no difference among the means of the three groups. How would you carry-out this test and what is the distribution of the proposed test statistic?

   d) Provide the least-squares estimates $\hat{\beta}$ algebraically.

   e) Provide an expression for Var-Cov matrix for $\hat{\beta}$.

3. For the experiment in problem #2 above, express the regression model as "cell-mean" model, that is $Y_{ij} = \mu_i + \epsilon_{ij}$. Where μ_i is the mean of the ith treatment group.

   a) Write the "cell-mean" model above in a matrix form as $Y = X\mu + \epsilon$, identify the design-matrix X and the parameters vector μ for the "cell-mean" model.

   b) State the null hypothesis of no difference among the means of the three groups.

c) Provide the least-squares estimates $\hat{\mu}$ algebraically.

4. A rehabilitation center researcher was interested in examining the relationship between physical ftness level prior to surgery of persons un-dergoing corrective knee surgery and time required in physical therapy until successful rehabilitation (in days). Patient records from the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The number of days required for successful recovery and fitness level (1=below average, 2=average and 3=above average) were recorded. The patient's age at the time of surgery was also recorded becuase previous studies have shown that younger patients tend to recover more qucikly. Data is given in HW3-prob4.dat file on Sakai.

   a) Fit the ANCOVA model. Based on this model fit, is there evidence that time to recovery differs across fitness levels after accounting for a patient's age? Please provide the hypotheses, test statistic, and p-value of the test used to adress this question.

   b) Using the model fit in part (a), provide the estimated regression model for each fitness level group.

   c) Estimate the overall slope beween days to recovery and patient age by removing fitness level from the model fit in part (a). How does this slope estimate compare to the slope estimates reported in part (b)?

   d) Compute the unadjusted mean days to recovery for each fitness level group. Using this numerical ouput, what factor level differences appear to be driving the global signal from fitness level if one exists?

   e) Compute the mean days to recovery for each fitness level adjusted for average patient age. Using this numerical ouput, what factor level difference appear to be driving the global signal from fitness level if one exists? Does your conclusion differ from the conclusion you reached in part (d)? If so, why does the discrepancy exist?

   f) Was it benefcial to inlcude patient age in this analysis when the primary goal was to assess the relationship between days to recovery and fitness level? Explain your reasoning.