APPM 4570/5570 — Final Exam — Spring 2017

HARD COPY due Thursday, May 11th, by 6:00pm, ECOT 218

You must solve all problems neatly and clearly. Correct answers with no supporting work may receive little or no credit. **Show all work, justify your answers, and interpret anything that seems important.**

THEORETICAL PORTION

100 points 4570, 140 points 5570

1. (30 points) A factory makes 2 types of batteries: long-lasting (most expensive), and short-term (least expensive). The long-lasting batteries have a lifetime that follows an exponential distribution with a mean of 1000 hours. The short-term batteries have a lifetime that follows a gamma distribution with a mean of 500 hours and a variance of 200 hours. The factory produces 4 long-term batteries for every short-term battery.

   (a) What are the values of the parameters for the short-term batteries?

   (b) What is the probability that a battery, picked at random from the factory, lasts at least 490 hours?

   (c) You are a QA inspector and have a battery in your hand that has lasted for 490 hours (and could possibly last longer). What is the probability it is a short-term battery?

2. (30 points) Let X be a random variable with pdf $f(x) = cx^2$, with $0 < x < 1$.

   (a) Solve for c in $f(x)$.

   (b) Let $Y = X^3$. What is $E(Y)$?

   (c) What is $Var(Y)$?

3. (40 points) Are more than 80% of Americans right handed? Researchers at a major university are trying to improve outcomes for left-handed individuals in the U.S. As a start, they sample 500 people in major cities across the country and ask if they are right- or left-handed. Among their sample, 419 people were right-handed.

   (a) Write out the hypothesis that is being tested.

   (b) What is the p-value for your test? Be sure to show your work.

   (c) What is the conclusion for your test? Be sure to interpret in terms of the original problem.

4. **APPM 5570 only** (40 points): Let $Y_i = \beta x_i + \epsilon_i$ be a linear regression model, with $\epsilon_i \sim \mathcal{N}(0, 5)$ (i.e., σ^2 is known).

   (a) Find the equation for the least squares estimate, $\hat{\beta}$, for β.

   (b) Show that your estimator is unbiased.

   (c) Find the variance for your estimator.

   (d) State the distribution of your estimator (provide both the type of distribution and the parameter values).

   (e) You have a data set of size 200. Summary values for your data are:

   $$\sum x_i = 200.34, \sum y_i = -40.36, \sum x_i y_i = -147.37, \sum x_i^2 = 603.26, \sum y_i^2 = 1,055.89.$$

   Based on the work you have done in the previous portions of this problem, is there significant evidence that the effect of x on Y is smaller than -0.2?

COMPUTATIONAL PORTION

150 points 4570, 170 points 5570

1. **Linear Regression** (100 points 4570, 120 points 5570)
   You have just been hired by a leading cereal company as chief statistician! You feel ready for the challenge, since you studied so hard in your statistical methods class. The company wants to make a healthy cereal that will also sell well, and your goal as the statistician is to determine which cereal variables lead to better FDA "health rankings".

   To assess this, you are given a data set that contains 77 different cereals (this is found in the "cereals.txt" file). For each cereal, these measurements (per 2 cups of cereal) are provided:

   mfr: Cereal manufacturer (G: General Mills, K: Kellogs, N: Nabisco, O: Other)

   calories: Calories per serving

   protein: Grams of protein per serving

   fat: Grams of fat per serving

   sodium: Mg of sodium per serving

   fiber: Grams of dietary fiber per serving

   carbo: Grams of complex carbohydrates per serving

   sugars: Grams of sugar per serving

   potass: Mg of potassium per serving

   vitamins: Percentage of total daily vitamins and minerals recommended by the FDA

   shelf: Display shelf of the cereal (levels range from 1 to 3, counting up from the floor)

   rating: Rating of the cereal by the FDA.

   Keep in mind that the head of the cereal company struggles with interpreting linear model output, and wants answers in terms of things that are easily read and understood. Therefore, when analyzing your models, be sure your answers are friendly for a general audience, but include enough technical information that your statistics professor believes you know what you're talking about.

   (a) Make a histogram of your outcome, which is "rating", and compare it to a histogram of "log(rating)". Which one looks more suitable for a linear model? Transform your data appropriately. (HINT: You need to log-transform your outcome before going any further!)

   (b) Make a scatterplot to examine the different predictor variables. Do any of them look correlated?

   (c) Using matrix notation, write out the *full* model for the first and last 5 observations (10 total) from the data. (Here, the term "full model" refers to the model that includes all covariates.) **Be sure to give your variables (βs, Xs) meaningful names.** For variables that are factors, be sure to include a column in X for each factor.

   (d) Fit the full model using `lm()` in R, and answer the following questions:

      i. Write out your full estimated model.

      ii. What is the average rating for a General Mills cereal that sits on the bottom shelf and contains the minimum number (among the cereals in the data set) of: calories, protein, fat, sodium, fiber, carbohydrates, sugars, potassium, and vitamins?

      iii. Does the manufacturer appear to have a statistically significant impact on the FDA rating? Do you think this is appropriate?

      iv. Does the shelf the cereal sits on appear to have a statistically significant impact on the FDA rating? Do you think this is appropriate?

   (e) Some of the covariates from the full model are not significant and are perhaps illogical in assessing FDA rankings. Create a smaller model that only includes these variables: calories, protein, fat, sodium, fiber, carbohydrates, sugars, and vitamins. Fit this model in R and answer these questions:

      i. Write out the estimated smaller model.

      ii. What is the average rating for a cereal that contains the minimum number (among the cereals in the data set) of: calories, protein, fat, sodium, fiber, carbohydrates, sugars, and vitamins? How does this compare to your larger model? Does this make sense?

      iii. How does the average rating change with a 1 gram increase in fat?

      iv. What percent of the variation in ratings is explained by your model?

(f) Perform a hypothesis test to determine if any of the predictors removed from the full model to create the smaller model are significant predictors of the rating. Provide the hypothesis, perform the test, and state the conclusions using your p-value. Be sure to provide your answer in terms of the original problem.

(g) Compare the percent of variation in ratings explained by your smaller and larger models. Which model do you think is better and why?

(h) Check that your model satisfies our modeling assumptions and discuss.

(i) **APPM 5570 only:** Use an automated routine in R (such as forward or backward stepwise regression, AIC or BIC selection criteria, etc) to find an "optimal" model:

i. How does your automated method work? Discuss the advantages and drawbacks of your automated method (be thorough).

ii. Does this model differ from the smaller model you have already created? Which model do you prefer?

2. **Regression Power Calculations** (50 points 4570/5570)
An education research specialist is working with a professor who teaches four pre-calculus classes at CU Boulder to test their hypothesis. Currently, pre-calculus test scores (which are normally distributed) are described with the following regression model:

$$Y_i = 60 + 20x_i + \epsilon_i, \tag{1}$$

where $\epsilon_i \sim N(0, 20^2)$, and

- Y_i is the score of midterm 1 for the i^{th} student
- x_i is an indicator variable that equals 1 if student i took pre-calculus in high school

The researcher is investigating if "active group work" improves test scores, and hypothesizes that this intervention will improve scores an average of at least 10 points.

(a) What students are in the baseline group in Equation (1)?

(b) Interpret the parameters in Equation (1).

(c) In a typical class, 75% of students have taken pre-calculus before. Additionally, half of the students will get the intervention and half will not, and the intervention must be evenly split percentage-wise among those who had pre-calculus and those who did not. Write down code for generating this X matrix in R for n students.

(d) Write down your psuedocode for calculating the power of this test for n students, assuming X does not change after it is generated.

(e) Describe your approach and psuedocode from (2d), in words.

(f) What sample size does the researcher need to detect their hypothesized difference of 10 at 80% power?