Homework 4

Due Thursday Aug 2, 2012 before 5PM

Please turn in a *physical* copy of the homework in my mailbox, located in the Statistics Department (4th floor, JMHH)

1. Download the "Wage" data set from the lecture on multiple regression. The series of questions below will examine various aspects about the age of the people in the U.S. workforce.

   (a) Suppose we want to predict the age of a male member of the labor force by looking at the (i) number of hours he works in a week, (ii) his martial status, (iii) his education year, (iv) his race, and (v) his salary. Fit a multiple regression model with those variables. What is the R^2? Is the model significant?

   (b) What is the predicted age of your instructor? He works 168 hours every week, has never married, been educated for 19 years, is Asian, and gets paid less than 50K. Provide a 95% CI and PI.
   Note: Make sure you only look at males in the data set before you fit this regression

   (c) *(Bonus Question):* What is the difference between adding the "Sex" variable in (a) as another predictor compared to only looking at a subset of the data (e.g. males in the data set)?

   (d) Are there any outliers in data? Use the residual plot, the H_{ii} values, and D_i to identify regression outliers, leverage points, and influential points.

   (e) Does the model satisfy the assumptions of regression? Check homoscedasticity, linearity, and normality using the appropriate plots. Transform the data, as needed, and refit the regression with the necessary modifications. What is the predicted age of your instructor? Provide a 95% CI and PI. How do these intervals differ from (a)?

   (f) Your instructor is considering whether to add capital gains and losses into the regression model in (e). He believes these two variables will significantly help, in liue with all the variables that are current in (e). Should he or should he not? Assume that the significance is $\alpha = 0.05$

   (g) Your instructor believes that the interaction between race and salary may play a crucial role in explaining the variation in age specified by the model in (e). Conduct a test to confirm or reject his belief. Assume that the significance is $\alpha = 0.01$.

2. Download the "Common Household Food" data set from the lecture on multiple regression. The series of questions below will examine various aspects about calories of common household food. *Make sure you take out the last missing observation!*

   (a) Fit a regression model that explains the calories in common household food. Use all the variables that are in the data set (except for the actual food name). What is the R^2 of this regression? Is this model significant?

(b) Check for any violations of the regression assumptions by using appropriate plots. Identify possible outliers. Refit the regression with the outliers removed and with the necessary adjustments (e.g. transformations). What is the R^2 of this data? Is this model significant?

(c) Your instructor believes that minerals are not significant predictors, controlling for other measurements of nutrition, in the model specified in (b). Conduct a hypothesis test and prove or disprove his claim.

(d) The data contains two measurements of Vitamin A, IU and RE. Your instructor wants to have only one measure of Vitamin A for the model specified in (b). Suggest which measure he should use by using the AIC, BIC, and Mallow's C_p. For each information criterion, which Vitamin A measurement is chosen as the best one for the mode?

(e) *(Bonus Question):* Why can't you use the F test to choose between the two measurements of Vitamin A?

(f) In R, design a K-fold cross validation procedure to decide which Vitamin A measurement to include. Conduct a 5-fold cross validation and compare the difference in PMSE between the model with IU and the model with RE. Which model has the lowest PMSE? Based on 5-fold CV, which measurement should the instructor use in the regression model?

(g) *(Bonus Question):* Your instructor wants to include exactly one nutritional measurement from each of the following categories

i. *Fats:* Fat (in grams), saturated fat (in grams), monounsaturated fat (in grams), polyunsaturated fat (in grams)
ii. *Protein:* Protein (in grams)
iii. *Carbohydrates:* Carbohydrates (in grams)
iv. *Cholesterol:* Cholesterol (in mg)
v. *Vitamins:* Vitamin A (in IU), Vitamin A (in RE), Vitamin B_1 (in mg), Vitamin B_2 (in mg), Vitamin B_3 (in mg), Vitamin C (in mg)
vi. *Minerals:* Calcium (mg), Phosphorus (mg), Iron (mg), Potassium (mg), Sodium (in mg)
vii. *Weight:* Weight (in grams)

in to a regression model about calories. Specifically, he wants

$$\text{Calories} \sim \beta_0 + \beta_1(\text{Fats}) + \beta_2(\text{Protein}) + \beta_3(\text{Carbs}) + \beta_4(\text{Vitamin}) + \beta_5(\text{Minerals})$$

Come up with a procedure to obtain the "best" model under this constraint. Here, "best" model is the model that has the most accurate prediction for a given value of X. You may use any procedure you wish (or you may come up with one of your own). What is your model?
Note: The student with the best prediction will get generous bonus points. If this student beats your instructor's model, he/she will definitely get at least an A- in the course.

Hint 1: I would check to see whether your final regression model follows the assumptions of regression. I woud also consider taking out possible outliers in your regression model.

Hint 2: To test your model's prediction power, consider the following food item in table

1. If your model doesn't predict a value close to 120 calories, I would suggest rebuilding your model.My model gave me 119.5 as the predicted value with a 99% prediction interval of (118,121)

3. Download the "Wine" data set from the lecture on multiple regression. In these questions, we'll fit a regression model between flavinoids and proline

   (a) Construct a polynomial regression that explains the variation in proline using flavanoids. Conduct an F test to determine which power terms in the polynomial you kept in your final model. What is your final model?

   (b) Describe the residual plot of the model you fitted in (a) and the QQ plot of the residuals. Based on the residual plot, is there any reason there are any violations of regression model assumptions?

   (c) Identify outliers in the model you fitted in (a).
   Hint: There is at least one obvious leverage/influential point!

   (d) What is the predicted value of proline if flavanoids is 2.5? What is the 95% PI?

Actual Calories	120
Fat	1.5 grams
Saturated fat	0 gram
Monounsaturated fat	0.5 grams
Polyunsaturated fat	0.5 grams
Protein	3 grams
Carbohydrates	26 grams
Cholesterol	0 mg
Vitamin A	1250 IU, 125 RE
Vitamin B_1, Vitamin B_2, Vitamin B_3	0 mg
Vitamin C	30 mg
Calcium, Phosphorus	0 mg
Iron	1.8 mg
Potassium	95 mg
Sodium	85mg
Weight	33g

Table 1: A food product for question 2g