Regression, Time Series, and Forecasting (STAT:4510)
Fall 2015
Homework 3 – due Wednesday, October 14, 2015

1. Mercury contamination of edible fish poses a health threat. Large-mouth bass were studied in Florida lakes to examine factors that influence the amount of mercury contamination. Water samples were collected and anlyzed from 38 lakes. Samples of three-year old fish were taken from each lake and mercury concentration in their muscle tissue was measured. Standardized mercury levels were then calculated for each lake. The data on the standardized mercury levels (Y) in ppm, and the alkalinity (X_1), calcium (X_2) and pH of the water (X_3) are in the file MERCURY.csv (given on ICON).

   (a) Make a scatterplot matrix for the four variables. It should be clear that from the plots that there is curvature in the relationships between Y and X_1 and between Y andX_2.

   (b) Make a scatterplot matrix of $log(Y)$, $log(X_1)$, and $log(X_2)$. Do the plots of $log(Y)$ vs. $log(X_1)$ and $log(Y)$ vs. $log(X_2)$ indicate linear relationships? Note: Use natural logs.

   (c) Fit the model $log(Y) = \beta_0 + \beta_1 log(X_1) + \beta_2 log(X_2) + \beta_3 X_3 + \varepsilon$.

   (d) Interpret $\widehat{\beta}_3$, the parameter estimate of β_3, in terms of standardized mercury level. Do not use log in your interpretation; that is, do not say "The logarithm of the standardized mercury level ..."

2. The file STOCK.csv gives data for 19 stocks on the following variables: profit margin (PM, X_1), growth rate (GR, X_2), type of industry (INDUSTRY, X_3), and price-to-earnings (P/E) ratio (PE, Y). The type of industry is coded as 1 = oil, 2= drug/healthcare, 3 = computer/electronics.

   (a) Fit a linear model to predict the P/E ratio as a function of the profit margin, growth rate, and the indicator variables for the type of industry, with the oil industry being the base level.

   (b) For the model in part (a), give an interpret the coefficient of the indicator variable that equals 1 if the type of industry is computer/electronics and 0 otherwise.

   (c) Supose we choose the drug/healthcare industry as the base level. **Using the results of your model from part (a)**, give the coefficient of all the indicator variables for this new model. (Hint: check your answer by fitting the new model)

3. A team of anthropologists and nutrition experts investigated the influence of protein content in a diet on the relationship between AGE and height (HT) for New Guinean children. Use the data given in the file DIET.csv (given on ICON) for a hypothetical sample of children with protein-rich and protein-poor diets, respectively.

   (a) Determine the least-squares line of HT (Y) on AGE (X) separately for Protein-Rich Diet (DIET="Rich") and for Protein-Poor Diet (DIET="Poor").

   (b) Produce a single scatterplot of the data and the regression lines. Use different plotting symbols for the two diet groups.

   (c) Test H_0: "The slopes are the same for the populations of high and poor diets being sampled," versus H_A: "The slopes are different."

   (d) Test H_0: "The intercepts are the same for the populations of high and poor diets being sampled," versus H_A: "The intercepts are different."

   (e) Test H_0: "The lines coincide for the populations of high and poor diets being sampled," versus H_A: "The lines do not coincide."

4. Use the data from Problem 3.

   (a) State the single regression model that incorporates the models for both rich and poor diet groups.

   (b) Fit the model you described in part (a).

   (c) Test for parallelism using this model.

   (d) Test for coincidence of the two lines.

   (e) Compare your results in parts (b), (c), and (d) with your results from Problem 3.