Statistics 512: Homework#7
Due March 14, 2014 BEFORE CLASS

1. Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data include the invasive measure of systolic pulmonary arterial pressure (Y) and three potential noninvasive predictor variables. Two were obtained by using radionuclide imaging – emptying rate of blood into the pumping chamber of the heart (X_1) and ejection rate of blood pumped out of the heart into the lungs (X_2) – and the third predictor variable measures a blood gas (X_3). The data are in `CH09PR13.DAT` with the columns for Y, X_1, X_2, and X_3 in order. Here we evaluate the regression model containing first-order terms for X_1 and X_2 and the cross-product term X_1X_2.

   (a) Obtain the residuals and plot them separately against $\hat{Y}$ and each of the three predictor variables. On the basis of these plots, should any further modifications of the regression model be attempted?

   (b) Take a normal probability plot of the residuals. Does the normality assumption appear to be reasonable here?

   (c) Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

   (d) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with $\alpha = 0.05$. State the decision rule and conclusion.

   (e) Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying X observations. Also prepare separate dot plots for each of the three predictor variables. Are there any consistent findings in the dot plots? Should they be consistent? Discuss.

   (f) Case 3, 8, and 15 are moderately far outlying with respect to their X vlaues, and case 7 is relatively far outlying with respect to its Y value. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to assess their influence. What do you conclude?

2. A group of high-technology companies agreed to share employee salary information in an effort to establish salary ranges for technical positions in research and development. Data obtained for each employee included current salary (Y), a coded variable indicating highest academic degree obtained (1=bachelor's degree, 2=master's degree, 3=doctoral degree), years of experience since last degree (X_3), and the number of persons currently supervised (X_4). The data are in `CH11PR08.DAT` with the columns for Y, academic degree, X_3, and X_4 in order.

   (a) Create two indicator variables for highest degree attained:

Degree	X_1	X_2
Bachelor's	0	0
Master's	1	0
Doctoral	0	1

(b) Regress Y on X_1, X_2, X_3, and X_4, using a first-order model and ordinary least squares, obtain the residuals, and plot them against $\hat{Y}$. What does the residual plot suggest?

(c) Divide the cases into two groups, placing the 33 cases with the smallest fitted values $\hat{Y}_i$ into group 1 and the other 32 cases into group 2. Calculate the mean squared errors of the two groups separately and compare them. What do they suggest?

(d) Plot the absolute residuals against X_3 and against X_4. What do these plots suggest about the relation between the standard deviation of the error term and X_3 and X_4?

(e) Estimate the standard deviation function by regressing the absolute residuals against X_3 and X_4 in first-order form, and then calculate the estimated weight for each case using (11.16a).

(f) Using the estimated weights, obtain the weighted least squares fit of the regression model. Are the weighted least squares estimates of the regression coefficients similar to the ones obtained with ordinary least squares in part (b)?

(g) Compare the estimated standard deviations of the weighted least squares coefficient estimates in part (f) with those for the ordinary least squares estimates in part (b). What do you find?

(h) Iterate the steps in parts (e) and (f) one more time. Is there a substantial change in the estimated regression coefficients? If so, what should you do?