STATISTICS 500 – Fall 2009

Homework 8, handed out Friday, 23 Oct 2009
on campus Friday, 30 Oct 2009, in lecture (11 am)
or by e-mailto Chuanlong, dclong@iastate.edu, no later than noon.
off campus Monday, 2 Nov 2009, by 4 pm to Nicole Rembert, email: rembeall@iastate.edu or
FAX: 515-294-4040 (please include cover page with Stat 500 / Nicole Rembert).

1. **Multiple regression and tests** — linerboard production
The file mill.txt contains production information about linerboard, a paper product. The amount of linerboard produced for each of 25 months is recorded (PRODUCT), along with the cost of raw materials (RAWMAT), energy usage in btus (ENERGY), mill depreciation (DEPREC) and labor costs (LABOR).

   In economics, a Cobb-Douglas production function has the form

   $$PRODUCT = \beta_o (RAWMAT)^{\beta_1} (ENERGY)^{\beta_2} (DEPREC)^{\beta_3} (LABOR)^{\beta_4} \times (ERROR).$$

   One hypothesis of particular interest is $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$. This is the hypothesis of constant returns to scale – if the hypothesis is true, then multiplying all of the predictors (inputs) by a constant would lead to the response (output) being multiplied by the same constant.

   Fit the Cobb-Douglas function to these data. Ask if you don't know or don't remember how to convert functions like this to a multiple linear regression. Then, answer the following questions.

   (a) Test the hypothesis that $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$. Report the test statistic and p-value; write an appropriate conclusion.

   (b) Test the hypothesis that $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ in the Cobb-Douglas model. Again, report the test statistic, a p-value and write an appropriate conclusion.

   (c) Estimate $\beta_1 + \beta_2 + \beta_3 + \beta_4$ and its s.e. Calculate a 95% c.i. for the sum.

   (d) A friend claims there is a simpler way to compute the s.e. of $\beta_1 + \beta_2 + \beta_3 + \beta_4$. Since the PROC REG output includes s.e.'s for each β_i, you could compute $\sqrt{se^2_{\beta_1} + se^2_{\beta_2} + se^2_{\beta_3} + se^2_{\beta_4}}$. This value is 0.182 for these data. Explain why this is not the s.e. of $\beta_1 + \beta_2 + \beta_3 + \beta_4$.

2. **Added variable plots** —
The data in avp.txt were constructed to make some points about assessing the form of f(X) in a multiple regression. The data set has 3 variables, X1, X2 and X3 that are to be used to predict the response Y. All my questions based on the model:

   $$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \varepsilon_i$$

   The investigators who collected these data are especially concerned about lack of this multiple regression.

   (a) Fit the regression and plot the residuals against each of the X variables. Is there any indication of lack of fit? Explain why or why not.

   (b) Fit a full quadratic (i.e. including $X1^2$, $X2^2$, $X3^2$, X1X2, X1X3, and X2X3 in the model). Use the results from this model and the original regression to test for lack of fit. Is there any evidence of lack of fit?

   (c) Fit a lowess non-parametric regression using a smoothing parameter of 0.6. Use the results from this model and the original regression to test for lack of fit. Is there any evidence of lack of fit?

(d) Examine added variable plots (aka. partial regression residual plots) for each variable? Is there any indication of lack of fit? Explain your answer.

3. **Case diagnostics and more** — Website development
The data in website.txt are from an observational study of productivity of developers at a website development company. The eventual goal of the analysis is to "determine which variables have the greatest impact on the number of websites delivered". Consider the model

$$DELIVER_i = \beta_0 + \beta_1 BACKLOG_i + \beta_2 EXPERIENCE_i + \beta_3 PROCESS_i + \beta_4 YEAR_i + \epsilon_i$$

Fit this regression model, then consider the following questions. Please consider to be a medium size sample.

(a) Consider each point. Any concerns about regression outliers? If so, list the points (by id #) that are a concern and briefly explain why.

(b) Do any points raise concerns about unusually large influence on the fitted values? If so, list the id #'s that are a concern and briefly explain why.

(c) Since the objective here is to examine regression coefficients, do any points have unusually large influence on the estimated regression coefficients? Again, if so, list the points (by id #) that are a concern and briefly explain why.

(d) Are there any concerns with multicollinearity for any variables? Explain why or why not.

(e) Plot residuals vs predicted values. Any issues that concern you? Explain why or why not.

(f) The CEO of this company reminds you that she hired you to "determine which variables have the greatest impact on the number of websites delivered". Use standardized regression coefficients to answer her question.

(g) When you explain the standardized regression coefficients, the clients are confused because the s.d. of PROCESS is 0.48. They don't understand what an 0.48 change in PROCESS means when PROCESS is either 0 or 1. After inspection of the X variables and discussion with the client, you agree that a change of 12 backlog orders is comparable to a change to 9 months experience, a change from process=0 to process=1, or a change from year =2001 to year = 2002. Which variable (or variables) have the largest impact on productivity?