Linear Regression Models
Statistics GR5205/GU4205 — Fall 2020

Homework 5

The following problems are due on Monday, Nov 23th, 11:59pm.

1. Continue with the *SENIC Project* data, specifically the regression of Y = number of nurses on X = available facilities and services.

   (a) Examine (but don't submit) boxplots and histograms for both variables. Does either seem to be a good candidate for a data transformation? Explain (you may also wish to refer back to the scatterplot).

   (b) Prepare an inverse response plot for the simple regression of `Nurses` on `AFS`. What response transformation is suggested by this method?

   What response transformation is suggested by the Box-Cox method? Answer with reference to a plot of $(-n/2)\log(SSE(\lambda)/n)$, ie the *log-Likelihood*, versus λ for various choices of λ.

   (c) Select a power transformation λ and fit the simple linear regression model

   $$Y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

   where

   $$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n \text{ are iid } N(0, \sigma^2)$$

   Interpret the value of b_1, the least squares estimate of β_1.

   (d) Prepare a scatterplot of residuals versus fitted values, and a normal probability plot of the residuals. Do the data seem to reasonably conform to the model assumptions?

   (e) Is it possible to conduct an F-test comparing the model in part (c) to the model you fit in problem 1(b) above? If so conduct the test; if not explain why.

   (f) Obtain separate 95% prediction intervals for the number of nurses at two hospitals, one with an AFS percentage of 30 and one with an AFS percentage of 60. How do the resulting intervals compare to those you computed for HW 4 problem 1(d)? Which intervals would you report to the client? Why?

2. (Project 9.25 in KNN) Consider again the *SENIC Project* data, this time regressing $Y =$ (Length of stay)$^{-1}$ on the predictor variables

Variable	Description
X_1	Average age of patients
X_2	Infection risk
X_3	log(Routine culturing ratio)
X_4	Routine chest X-ray ratio
X_5	log(Number of beds)
X_6	log(Average daily census)
X_7	log(Number of nurses)
X_8	Available facilities and services

(a) Use the Box-Cox methodology to justify the choice of response and predictor transformations suggested in the variable definitions above.

(b) Examine (but do not submit) a scatterplot matrix of response and all predictors. Which batch of predictor variables are most highly correlated?

(c) Run a forward selection algorithm, starting with the intercept-only model, and taking all eight predictor variables as the scope. Which variables are included in the mean function suggested by the forward selection algorithm?

(d) Run the backward elimination routine on the mean function that includes all eight predictor variables. Does this algorithm choose the same mean function as forward selection?

(e) Present and interpret a final fitted model of your choosing, that you feel provides the best available aid to our understanding of how the average length of hospital stay is related to the hospital characteristics encompassed by the variables in the *SENIC* dataset.

3. (Problem 9.10 in KNN) A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) for the 25 applicants were as given in the data file `job_proficiency.txt`, available in the CourseWorks Data folder.

   (a) Prepare a set of adjacent box plots for the test scores of the four newly developed aptitude tests (that is, all four box plots on one set of axes). Are there any noteworthy features in these plots? Comment.

   (b) Obtain the scatterplot matrix for response and predictor variables combined. What do the scatterplots suggest abut the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

   (c) Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained? Explain.

4. (Problem 9.18 in KNN) Continue with the *Job proficiency* data from the previous exercise.

   (a) Run the backward elimination algorithm using *AIC* as the model selection criterion. What mean function is selected by the algorithm?

   (b) Run a forward selection routine, again choosing a "best" model based on *AIC*. What mean function does this algorithm suggest?

   (c) Repeat parts (a) and (b) using the Bayesian criterion *SBC* instead of *AIC*. Does your conclusion about the "best" mean function change? Explain.