Homework 5
STAT W4315: Linear Regression Models
Multiple Linear Regression Model

DUE: **Thursday, April 28, 12:00 noon**

(1) Please sign your homework with your name and UNI number.

(2) Homework must be submitted into the Statistics Homework Boxes room 904 on the 9th floor of SSW building.

(3) Homework is due **Thursday, April 28, 12:00 noon**.

(4) No late homework, under any circumstances, will be accepted.

(5) At the end of semester, one of your lowest homework scores will be dropped before the final grade is calculated.

Problem 1 (35p) (Problem 9.10 b,c, Problem 9.11 a & Problem 9.18 a,b)

A personnel officer in a governmental agency administrated four newly developed aptitude test 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 employees were as follows

Y	X_1	X_2	X_3	X_4
88	86	110	100	87
80	62	97	99	100
96	110	107	103	103
...	...	...	...	...
83	91	129	97	83

(see file "Homework 5 data Problem1and2.txt" for a complete set of data)

(a)(5p) Obtain the scatter plot matrix, and the correlation matrix of the X variables. What do the scatter plots suggest about the nature of the functional relationship between the response variable Y and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

(b)(5p) Fit the multiple regression function containing all four predictor variables as first-order terms. Does it appear that all predictor variables should be retained?

(c)(10p) Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to the adjusted R^2 criterion.

(d)(10p) Using forward stepwise regression find the best subset of predictor variables to predict job proficiency. Use α limits of 0.05 and 0.10 for adding or deleting a variable, respectively (see Lecture 14, slides 11-12).

(e)(5p) How does the best subset according to forward stepwise regression compare with the best subset according to the adjusted R^2 criterion from (c) above?

Problem 2 (65p) (Problem 10.19 a,b,c,d,e (first part),f,g

The subset model from Problem 1 containing only first-order terms in X_1 and X_3 is to be evaulated in detail.

(a)(5p) Obtain the residuals and plot them separately against $\widehat{Y}$, each of the four predictor variables, and the cross-product term X_1X_3. On the basis of these plots, should any modifications in the regression model be investigated?

(b)(10p) Prepare separate added-variable plots against $e\left(X_1 \mid X_3\right)$ and $e\left(X_3 \mid X_1\right)$. Do these plots suggest that any modifications in the model form are warranted?

(c)(10p) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumptions, using $\alpha = 0.1$. What do you conclude?

(d)(10p) Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test (a t-test with $t\left(1-\alpha/(2N), N-P-1\right)$ critical value) with $\alpha = 0.05$. State the decision rule and conclusion.

(e)(10p) Obtain the diagonal elements of the hat matrix. Using the rule of thumb (X_i is outlying case with regard to the X values, if $h_{ii} > 2P/N$), identify any outlying X observations.

(f)(10p) Case 7 and 18 appear to be moderately outlying with respect to their X values, and case 16 is reasonably far outlying with respect to its Y value. Obtain $DFFITS$, $DFBETAS$, and Cook's distance values for these cases to asses their influence. What do you conclude?

(g)(10p) Obtain the variance inflation factors. What do they indicate?