Stat 615/415 (Regression) Homework # 9

Variable Selection and Model Building (chap. 9)

1. (**9.1**) A speaker stated: "In well-designed experiments involving quantitative explanatory variables, a procedure for reducing the number of explanatory variables after the data are obtained is not necessary." Do you agree? Discuss.

2. (**9.5**) In forward stepwise regression, what advantage is there in using a relatively small α-to-enter value for adding variables? What advantage is there in using a larger α-to-enter value?

3 (Don't Do) Two regression models are compared. The full model uses p independent variables $X_1, \ldots, X_p$. The reduced model uses only the first q variables $X_1, \ldots, X_q$, where $q < p$. We can choose the model with the higher adjusted R^2, or we can test significance of added variables $X_{q+1}, \ldots, X_p$ with a partial F-test. These methods are related!

   Show that the full model has a higher adjusted R^2 if and only if the F-statistic for testing $H_0 : \beta_{q+1} = \ldots = \beta_p = 0$ exceeds 1.

   Hint: Write explicit formulae for R^2_{adj} *and* F *and express both of them in terms of* $\frac{SSErr^{(Red)}}{SSErr^{(Full)}}$, *the ratio of error sums of squares from the two models.*

4. (Continuing 6.27 from an earlier homework) In a small-scale regression study, the following data were obtained,

Y	$X1$	$X2$
42.0	7.0	33.0
33.0	4.0	41.0
75.0	16.0	7.0
28.0	3.0	49.0
91.0	21.0	5.0
55.0	8.0	31.0

   Select the best regression equation using different model selection methods.

5. (**9.10–9.11, 9.18, 9.21–9.22**) 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, and scores of the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) were recorded.

   The resulting **Job Proficiency** data set is available on our Blackboard in "Data sets" and on the next page of this homework assignment.

   (a) Obtain the scatter plot matrix of these data. What do the scatter plots suggest about the nature of the functional relationship between the response variable and each of the predictor variables? Do you notice any serious multicollinearity problems?

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

(c) Using only first-order terms for the predictor variables in the pool of potential X-variables, find the best regression models according to different criteria - adjusted R^2, C_p, and BIC.

(d) Using forward stepwise selection, find the best subset of predictor variables to predict job proficiency. Use the α-to-enter limit of 0.05.

(e) Repeat the previous question using the backward elimination method and the α-to-remove limit of 0.10.

(f) To assess and compare internally the predictive ability of our models, split the data into training and testing subsets and estimate the mean squared prediction error MSPE for all regression models identified in (b–e).

(g) To assess and compare externally the validity of our models, 25 a1dditional applicants for entry level clerical positions were similarly tested and hired. Their data are below, in the table on the right. Use these data as the testing set and estimate MSPE for all regression models identified in (b–e).

Original Job Proficiency data

Y	X_1	X_2	X_3	X_4
88.0	86.0	110.0	100.0	87.0
80.0	62.0	97.0	99.0	100.0
96.0	110.0	107.0	103.0	103.0
76.0	101.0	117.0	93.0	95.0
80.0	100.0	101.0	95.0	88.0
73.0	78.0	85.0	95.0	84.0
58.0	120.0	77.0	80.0	74.0
116.0	105.0	122.0	116.0	102.0
104.0	112.0	119.0	106.0	105.0
99.0	120.0	89.0	105.0	97.0
64.0	87.0	81.0	90.0	88.0
126.0	133.0	120.0	113.0	108.0
94.0	140.0	121.0	96.0	89.0
71.0	84.0	113.0	98.0	78.0
111.0	106.0	102.0	109.0	109.0
109.0	109.0	129.0	102.0	108.0
100.0	104.0	83.0	100.0	102.0
127.0	150.0	118.0	107.0	110.0
99.0	98.0	125.0	108.0	95.0
82.0	120.0	94.0	95.0	90.0
67.0	74.0	121.0	91.0	85.0
109.0	96.0	114.0	114.0	103.0
78.0	104.0	73.0	93.0	80.0
115.0	94.0	121.0	115.0	104.0
83.0	91.0	129.0	97.0	83.0

Additional Data for (f $-$ g)

Y	X_1	X_2	X_3	X_4
58.0	65.0	109.0	88.0	84.0
92.0	85.0	90.0	104.0	98.0
71.0	93.0	73.0	91.0	82.0
77.0	95.0	57.0	95.0	85.0
92.0	102.0	139.0	101.0	92.0
66.0	63.0	101.0	93.0	84.0
61.0	81.0	129.0	88.0	76.0
57.0	111.0	102.0	83.0	72.0
66.0	67.0	98.0	98.0	84.0
75.0	91.0	111.0	96.0	84.0
98.0	128.0	99.0	98.0	89.0
100.0	116.0	103.0	103.0	103.0
67.0	105.0	102.0	88.0	83.0
111.0	99.0	132.0	109.0	105.0
97.0	93.0	95.0	106.0	98.0
99.0	99.0	113.0	104.0	95.0
74.0	110.0	114.0	91.0	78.0
117.0	128.0	134.0	108.0	98.0
92.0	99.0	110.0	96.0	97.0
95.0	111.0	113.0	101.0	91.0
104.0	109.0	120.0	104.0	106.0
100.0	78.0	125.0	115.0	102.0
95.0	115.0	119.0	102.0	94.0
81.0	129.0	70.0	94.0	95.0
109.0	136.0	104.0	106.0	104