LAB 6: MODEL SELECTION AND DIAGNOSTIC[1]

Problem 1. Refer to Patient satisfaction Problem 6.15. The hospital administrator wishes to determine the best subset of predictor variables for predicting patient satisfaction.

a. Indicate which subset of predictor variables you would recommend as best for predicting patient satisfaction according to each of the following criteria: (1) $R^2_{a,p}$, (2) AIC_p, (3) C_p, (4) BIC_p. Support your recommendations with appropriate graphs.

b. Do the four criteria in part (a) identify the same best subset? Does this always happen?

(Option) c. Would forward stepwise regression have any advantages here as a screening procedure over the all-possible-regressions procedure?

Problem 2. Grocery retailer.

A large, national grocery retailer tracks productivity and costs of its facilities closely. Data below were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X1), the indirect costs of the total labor hours as a percentage (X2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X3), and the total labor hours (Y).

i:	1	2	3	. . .	50	51	52
X_{i1}:	305,657	328,476	317,164	. . .	290,455	411,750	292,087
X_{i2}:	7.17	6.20	4.61	. . .	7.99	7.83	7.77
X_{i3}:	0	0	0	. . .	0	0	0
Y_i:	4264	4496	4317	. . .	4499	4186	4342

a. Obtain the studentized deleted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with α = .05. State the decision rule and conclusion.

b. Obtain the diagonal element of the hat matrix. Identify any outlying X observations.

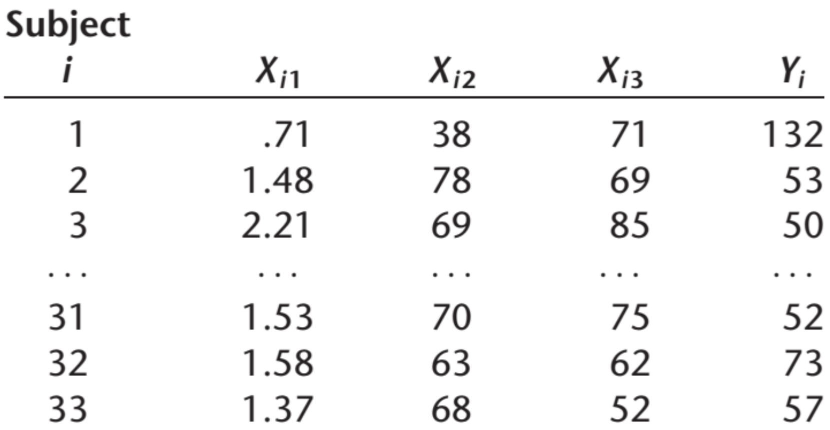
c. Management wishes to predict the total labor hours required to handle the next shipment containing X1 = 300,000 cases whose indirect costs of the total hours is X2 = 7.2 and X3 = 0 (no holiday in week). Construct a scatter plot of X2 against X1 and determine visually whether this prediction involves an extrapolation beyond the range of the data. Also, use (10.29) to determine whether an extrapolation is involved. Do your conclusions from the two methods agree?

[1] Reference: Chapters 9 and 10, Kutner's book.

d. Cases 16, 22, 43, and 48 appear to be outlying X observations, and cases 10, 32, 38, and 40 appear to be outlying Y observations. Obtain the DFFITS, DFBETAS, and Cook's distance values for each of these cases to assess their influence. What do you conclude?

e. Calculate Cook's distance Di for each case and prepare an index plot. Are any cases influential according to this measure?

Problem 3: Kidney function. Creatinine clearance (Y) is an important measure of kidney function, but is difficult to obtain in a clinical office setting because it requires 24-hour urine collection. To determine whether this measure can be predicted from some data that are easily available, a kidney specialist obtained the data that follow for 33 male subjects. The predictor variables are serum creatinine concentration (X1), age (X2), and weight (X3).



Subject i	X_{i1}	X_{i2}	X_{i3}	Y_i
1	.71	38	71	132
2	1.48	78	69	53
3	2.21	69	85	50
...	...	...	...	...
31	1.53	70	75	52
32	1.58	63	62	73
33	1.37	68	52	57

Adapted from W. J. Shih and S. Weisberg, "Assessing Influence in Multiple Linear Regression with Incomplete Data," *Technometrics* 28 (1986), pp. 231–40.

a. Fit the multiple regression function containing the three predictor variables as first-order terms. Obtain the variance inflation factors. Are there indications that serious multicollinearity problems exist here? Explain.

b. Obtain the residuals and plot them separately against $\hat{Y}$ and each of the predictor variables. Also prepare a normal probability plot of the residuals. Discuss.

c. What is added-variable plot? How is it used for? Prepare separate added-variable plots against e(X1|X2, X3), e(X2|X1, X3), and e(X3|X1, X2). Discuss.

Homework:

Problem 4. Refer to Real estate sales data set in Appendix C.7. Residential sales that occurred during the year 2002 were available from a city in the midwest. Data on 522 arms-length transactions include sales price, style, finished square feet, number of bedrooms, pool, lot size, year built, air conditioning, and whether or not the lot is adjacent to a highway. The city tax assessor was interested in predicting sales price based on the demographic variable information given above.

a) Select a random sample of 300 observations to use in the model-building data set. Develop a best subset model for predicting sales price. Justify your choice of model. Assess your model's ability to predict and discuss its use as a tool for predicting sales price.

b) Fit the regression model identified above to the validation data set. Compare the estimated regression coefficients and their estimated standard errors with those obtained in a). Also compare the error mean square and coefficients of multiple determination. Does the model fitted to the validation data set yield similar estimates as the model fitted to the model-building data set?

c) Calculate the mean squared prediction error (9.20) and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here?

Problem 5. P11. 9.10, 9.18, 9.22

-- END --