Regression Analysis (I)

Kutner's Applied Linear Statistical Models (5/E)

Chapter 9: Model Selection and Validation

Thursday 09:10-12:00, 資訊 140301

Han-Ming Wu

Department of Statistics, National Chengchi University

http://www.hmwu.idv.tw

9.1 Overview of Model-Building Process

A strategy for the building of a regression model:

1. Data collection and _____
2. Reduction of explanatory or _____ variables (for exploratory observational studies)
3. Model refinement and _____
4. Model _____

FIGURE 9.1
Strategy for Building a Regression Model.

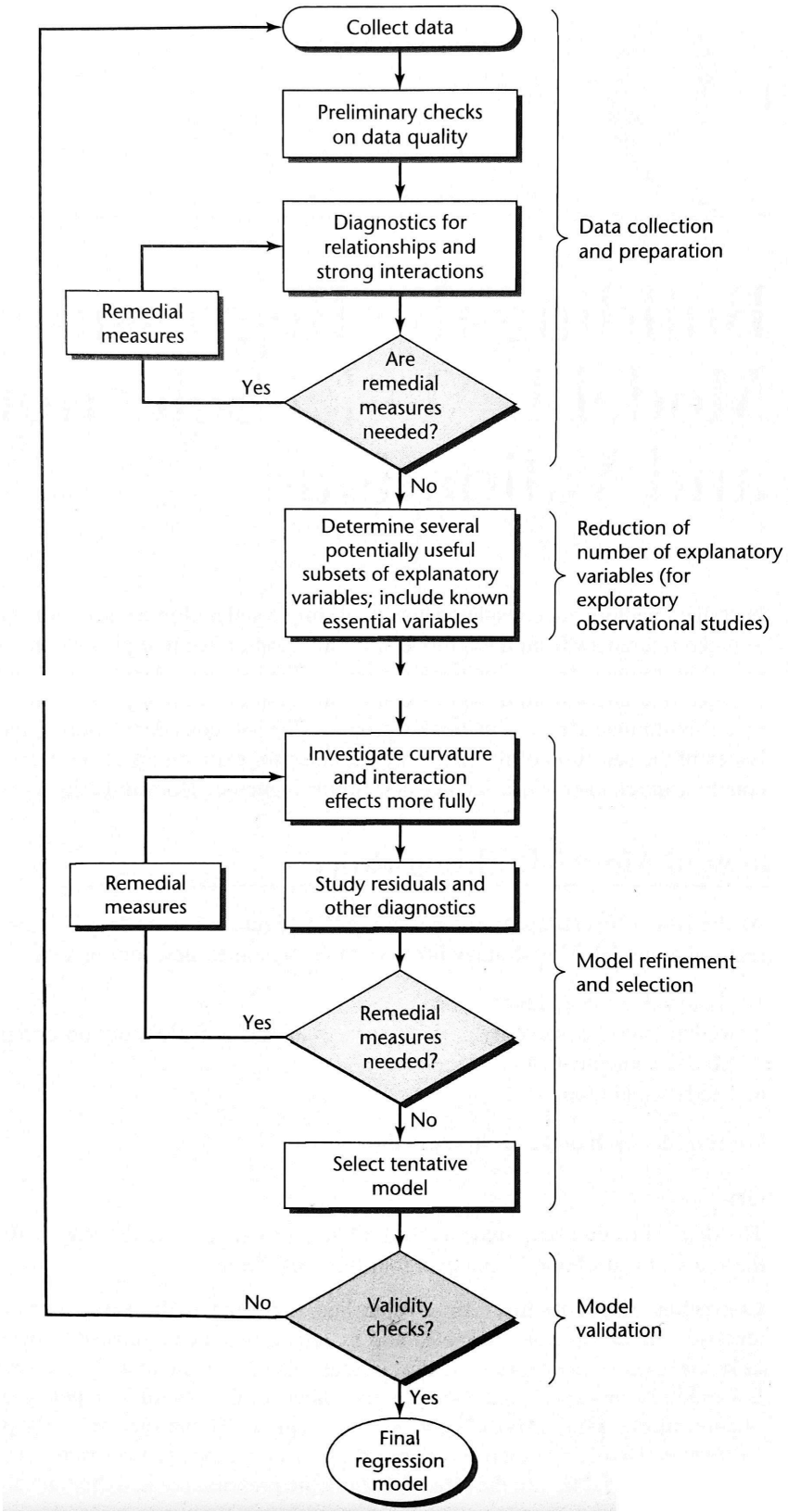


FIGURE 9.1
Strategy for Building a Regression Model.

9.2 Surgical Unit Example

1. A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. A random selection of 108 patients was available for analysis. From each patient record, the following information was extracted from the pre-operation evaluation:

X_1	blood clotting score (血栓分數)
X_2	prognostic index (預後指數)
X_3	enzyme function test score (酶功能)
X_4	liver function test score (肝功能)
X_5	age, in years
X_6	indicator variable for gender (0 = male, 1 =female)
X_7, X_8	indicator variables for history of alcohol use:
	None: $X_7 = 0, X_8 = 0$, Moderate: $X_7 = 1, X_8 = 0$,Severe:$X_7 = 0, X_8 = 1$

2. These constitute the pool of _____ or predictor variables for a predictive regression model.

3. (Table 9.1) The response variable Y is _____, which was ascertained in a follow-up study. A portion of the data on the potential predictor variables and the response variable is presented in Table 9.1. These data have already been _____ and properly _____ for errors.

TABLE 9.1 Potential Predictor Variables and Response Variable—Surgical Unit Example.

Case Number i	Blood-Clotting Score X_{i1}	Prognostic Index X_{i2}	Enzyme Test X_{i3}	Liver Test X_{i4}	Age X_{i5}	Gender X_{i6}	Alc. Use: Mod. X_{i7}	Alc. Use: Heavy X_{i8}	Survival Time Y_i	$Y_i' = \ln Y_i$
1	6.7	62	81	2.59	50	0	1	0	695	6.544
2	5.1	59	66	1.70	39	0	0	0	403	5.999
3	7.4	57	83	2.16	55	0	0	0	710	6.565
...	...	...	...	...	...	...	...	...	...	...
52	6.4	85	40	1.21	58	0	0	1	579	6.361
53	6.4	59	85	2.33	63	0	1	0	550	6.310
54	8.8	78	72	3.20	56	0	0	0	651	6.478

4. To illustrate the model-building procedures discussed in this and the next section, we will use only the *first four explanatory variables*. We will also use only the *first 54 of the 108 patients*.

5. Since the pool of predictor variables is small, a reasonably _____ of relationships and of possible strong interaction effects is possible at this stage of data preparation.

   (a) *Stem-and-leaf plots* for each of the predictor variables (not shown). These highlighted several cases as _____ with respect to the explanatory variables. The investigator was thereby alerted to examine later the _____ of these cases.

   (b) *A scatter plot matrix and the correlation matrix* (not shown)

6. *A first-order regression model* based on all predictor variables was fitted to serve as a starting point.

   (a) (Figure 9.2a) *A plot of residuals against predicted values* suggests that both _____ and _____ are apparent.

   (b) (Figure 9.2b) *the normal probability plot* suggests some _____ from normality.

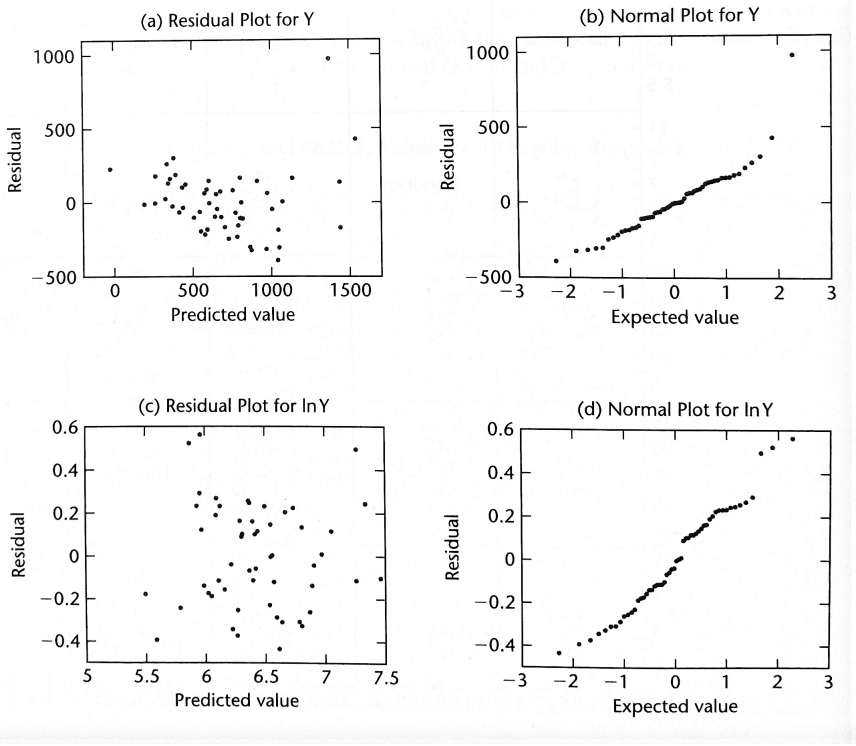


FIGURE 9.2 **Some Preliminary Residual Plots—Surgical Unit Example.**

7. *Transformation*: To make the distribution of the error terms more nearly normal and to see if the same transformation would also reduce the apparent curvature, the investigator examined the _____ transformation _____.

   (a) (Figure 9.2c) *A plot of residuals against fitted values* when Y' is regressed on all four predictor variables in a first-order model;

   (b) (Figure 9.2d) *The normal probability plot of residuals* for the transformed data shows that the distribution of the error terms is more _____.

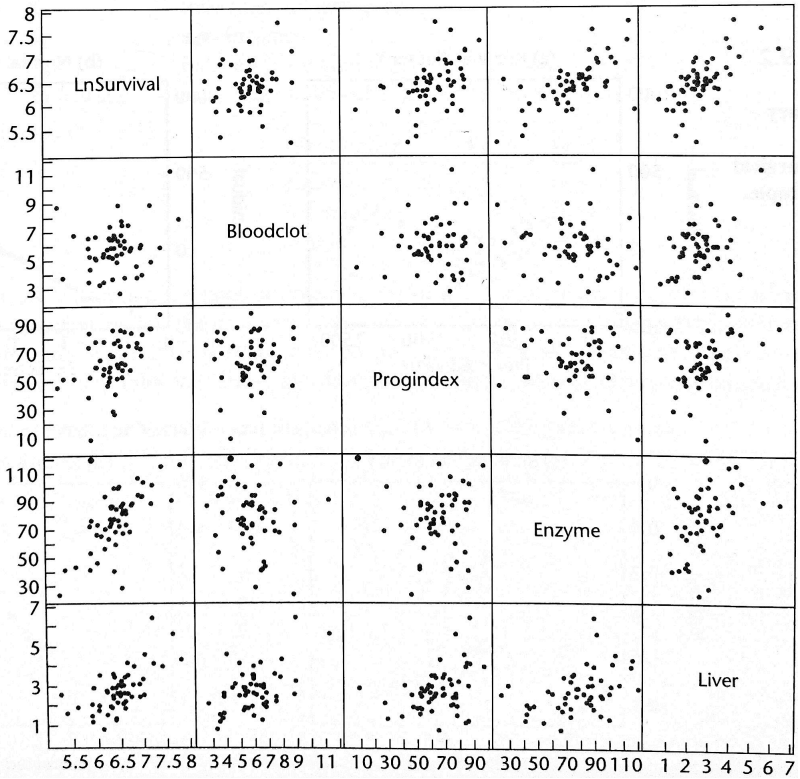
8. (Figure 9.3) *A scatter plot matrix and the correlation matrix* with the transformed Y variable.

FIGURE 9.3 JMP Scatter Plot Matrix and Correlation Matrix when Response Variable Is Y'—Surgical Unit Example.

Multivariate Correlations

	LnSurvival	Bloodclot	Progindex	Enzyme	Liver
LnSurvival	1.0000	0.2462	0.4699	0.6539	0.6493
Bloodclot	0.2462	1.0000	0.0901	−0.1496	0.5024
Progindex	0.4699	0.0901	1.0000	−0.0236	0.3690
Enzyme	0.6539	−0.1496	−0.0236	1.0000	0.4164
Liver	0.6493	0.5024	0.3690	0.4164	1.0000

Scatterplot Matrix



(a) Each of the predictor variables is _____ with Y', with X_3 and X_4 showing the highest degrees of association and X_1 the lowest.

(b) Show _____ among the potential predictor variables. In particular, X_4 has moderately high pairwise correlations with X_1, X_2, and X_3

9. Various _____ and _____ were obtained (not shown here).

10. On the basis of these analyses, the investigator concluded to use, at this stage of the model-building process, _____ as the response variable, to represent the predictor variables in linear terms, and not to include any interaction terms.

11. The next stage is to examine whether all of the _____ variables are needed or whether a subset of them is adequate.

9.3 Criteria for Model Selection

1. From any set of _____ predictors, _____ alternative models can be constructed. This calculation is based on the fact that each predictor can be either included or excluded from the model.

2. (Table 9.2) the _____ different possible subset models that can be formed from the pool of four X variables in The Surgical Unit Example.

TABLE 9.2 SSE_p, R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p, and $PRESS_p$ **Values for All Possible Regression Models—Surgical Unit Example.**

X Variables in Model	(1) p	(2) SSE_p	(3) R_p^2	(4) $R_{a,p}^2$	(5) C_p	(6) AIC_p	(7) SBC_p	(8) $PRESS_p$
None	1	12.808	0.000	0.000	151.498	−75.703	−73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	−77.079	−73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	−87.178	−83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	−103.827	−99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	−103.262	−99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	−88.162	−82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	−114.658	−108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	−102.067	−96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	−130.483	−124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	−107.324	−101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	−121.113	−115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	0.743	3.391	−146.161	−138.205	3.914
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	−105.748	−97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	−120.844	−112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	−138.023	−130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	−144.590	−134.645	4.069

3. _____ procedures, also known as subset selection or _____ procedures, have been developed to identify a small group of regression models that are _____ according to a specified criterion.

4. While many criteria for comparing the regression models have been developed, we will focus on six: _____.

5. We shall denote the number of potential X variables in the pool by _____. We assume throughout this chapter that all regression models contain an intercept term _____. Hence, the regression function containing all potential X variables contains _____ parameters, and the function with no X variables contains one parameter (β_0).

6. The number of X variables in a subset will be denoted by _____, as always, so that there are _____ parameters in the regression function for this subset of X variables. Thus, we have: $1 \leq p \leq P$.

7. We will assume that the number of observations exceeds the maximum number of potential parameters: _____.

R_p^2 or SSE_p Criterion

1. R_p^2 criterion calls for the use of the coefficient of _____:

$$R_p^2 = \_\_\_\_\_$$

2. R_p^2 indicates that there are p parameters, or _____ X variables, in the regression function on which R_p^2 is based.

3. The R_p^2 criterion is equivalent to using the error sum of squares _____ as the criterion (we again show the number of parameters in the regression model as a subscript).

4. The R_p^2 criterion is not intended to identify the subsets that maximize this criterion.

5. We know that R_p^2 can never decrease as _____ variables are included in the model. Hence, R_p^2 will be a _____ when _____ potential X variables are included in the regression model.

6. The intent in using the R_p^2 criterion is to find the point where _____ variables is not worthwhile because it leads to a very _____.

7. Example The Surgical Unit Example

   (a) (Table 9.2, column 3) the R_p^2 values were obtained from a series of computer runs.

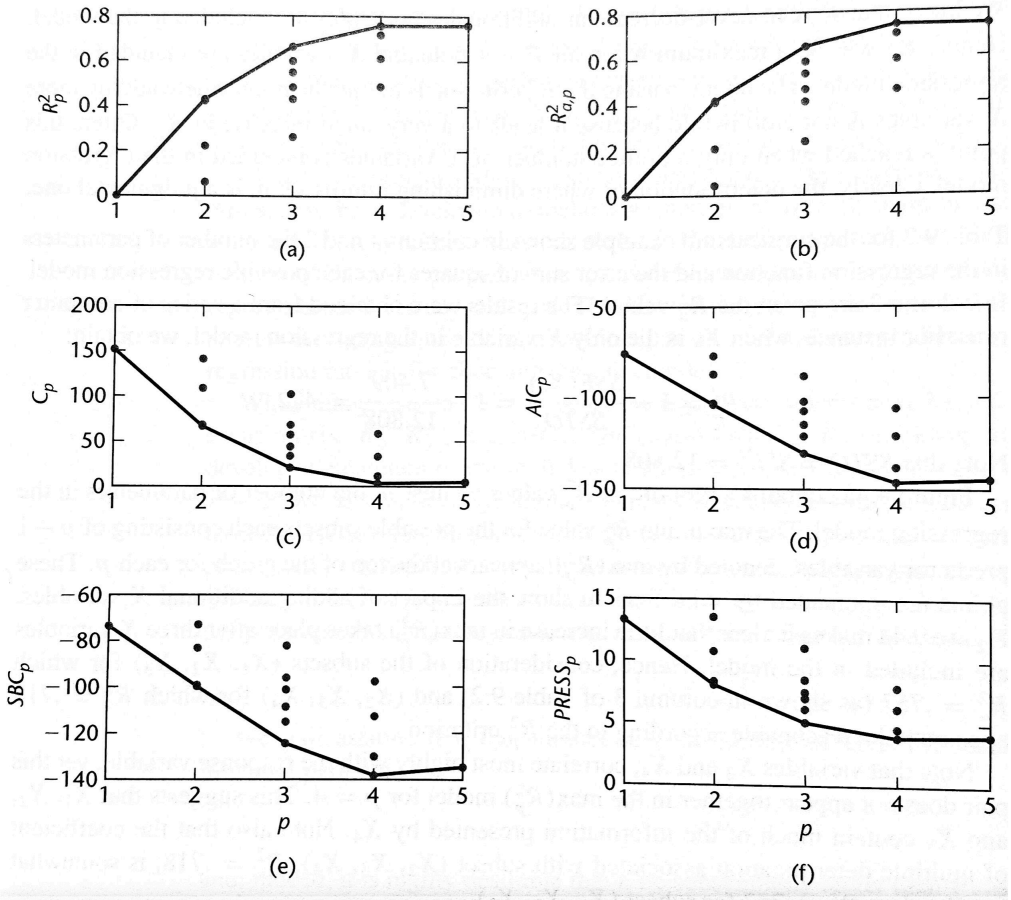
   (b) For instance, when X_4 is the only X variable in the regression model, we obtain:

   $$R_2^2 = 1 - \frac{SSE(X_4)}{SSTO} = \_\_\_\_\_$$

   Note that $SSTO = SSE_1 = 12.808$

   (c) (Figure 9.4a) a plot of the R_p^2 values against p, the number of parameters in the regression model.

FIGURE 9.4 Plot of Variables Selection Criteria—Surgical Unit Example.



(d) The maximum R_p^2 value for the possible subsets each consisting of $p-1$ predictor variables, denoted by _____, appears at the top of the graph for each p. These points are connected by solid lines to show the impact of _____.

(e) (Figure 9.4a) little increase in $\max(R_p)$ takes place after three X variables are included in the model.

(f) Hence, consideration of the subsets _____ for which $R_4^2 = 0.757$ (as shown in column 3 of Table 9.2) and _____ for which $R_4^2 = 0.718$ appears to be reasonable according to the R_p^2 criterion.

(g) Note that variables X_3 and X_4, correlate most _____ with the response variable, yet this pair does not appear together in the $\max(R_p^2)$ model for $p = 4$.

$R_{a,p}^2$ or MSE_p Criterion

1. Since R_p^2 does not take account of the _____ in the regression model and since $\max(R_p^2)$ can never decrease as p increases, the _____ of multiple determination $R_{a,p}^2$ in (6.42) has been suggested as an alternative criterion:

$$R_{a,p}^2 = \text{_____} \qquad (9.4)$$

2. It can be seeg from (9.4) that $R_{a,p}^2$ increases if and only if _____ decreases since $SSTO/(n-1)$ is fixed for the given Y observations. Hence, $R_{a,p}^2$ and MSE_p provide _____ information.

3. The largest $R_{a,p}^2$ for a given number of parameters in the model, $\max(R_{a,p}^2)$, can, indeed, _____.

4. Find a few subsets for which $R_{a,p}^2$ is at the _____ or so _____ the maximum that _____ more variables is not worthwhile.

5. Example The Surgical Unit Example

   (a) (Table 9.2, column 4). For instance, we have for the regression model containing only X_4:

$$R_{a,2}^2 = \text{_____}$$

(b) (Figure 9.4b) The story told by the $R^2_{a,p}$ plot in Figure 9.4b is _____ to that told by the R^2_p plot in Figure 9.4a.

(c) Consideration of the subsets _____ and _____ appears to be reasonable according to the $R^2_{a,p}$ criterion.

(d) Notice that _____ is maximized for subset _____, and that adding _____ to this subset − thus using all four predictors − decreases the criterion slightly: _____.

Mallows' C_p Criterion*

AIC_p and SBC_p Criteria

1. Two popular alternatives that also provide penalties for adding predictors are _____ _____ and _____.

2. We search for models that have small values of AIC_p, or SBC_p:

$$AIC_p = \_\_\_\_\_ \quad (9.14)$$

$$SBC_p = \_\_\_\_\_ \quad (9.15)$$

3. Notice that for both of these measures, the first term is $n \ln SSE_p$ which _____ as _____, The second term is _____ (for a given sample size n), and the third term _____ with the number of parameters, _____.

4. Models with _____ will do well by these criteria as long as the penalties − $2p$ for AIC_p and $(\ln n)p$ for SBC_p − are _____.

5. If _____ the penalty for SBC_p is larger than that for AIC_p.

6. Example The Surgical Unit Example

(a) (Table 9.2, columns 6 and 7) When X_4 is the only X variable in the regression model:

$$\begin{aligned} AIC_2 &= n \ln SSE_2 - n \ln n + 2p \\ &= \_\_\_\_\_ \\ SBC_2 &= n \ln SSE_2 - n \ln n + (\ln n)p \\ &= \_\_\_\_\_ \end{aligned}$$

(b) (Figures 9.4d, e) both of AIC_p and SBC_p criteria are minimized for subset _____.

$PRESS_p$ Criterion

1. The _____ criterion is a measure of how well the use of the _____ for a subset model can predict the _____ _____. The error sum of squares, _____, is also such a measure.

2. The $PRESS$ measure differs from SSE in that each fitted value Y_i for the $PRESS$ criterion is obtained by _____ from the data set, estimating the regression function for the subset model from the _____, and then using the fitted regression function to obtain the predicted value _____ for the ith case.

3. We use the notation _____ now for the fitted value to indicate, by the first subscript i, that it is a _____ for the ith case and, by the second subscript (i), that the ith case was _____ when the regression function was fitted.

4. The $PRESS$ prediction error for the ith case then is:

   _____ (9.16)

   and the $PRESS_p$ criterion is the sum of the squared prediction errors over all n cases:

   $$PRESS_p = \text{_____} \quad (9.17)$$

5. Models with _____ are considered good candidate models. The reason is that when the prediction errors $Y_i - \hat{Y}_{i(i)}$ are small, so are the squared prediction errors and the sum of the squared prediction errors.

6. Example The Surgical Unit Example

   (a) (Table 9.2, column 8)(Figure 9.4f) The message given by the $PRESS_p$ values in Table 9.2 and plot in Figure 9.4f is very _____ to that told by the other criteria.

(b) We find that subsets _____ and _____ have small $PRESS$ values;

(c) The set of all X variables (X_1, X_2, X_3, X_4) involves a slightly larger $PRESS$ value than subset (X_1, X_2, X_3).

(d) The subset (X_2, X_3, X_4) involves a $PRESS$ value of 4.597, which is moderately larger than the $PRESS$ value of 3.914 for subset (X_1, X_2, X_3).

9.4 Automatic Search Procedures for Model Selection

1. The number of possible models, _____, grows rapidly with the number of predictors.

2. A variety of _____ procedures have been developed, e.g., "best" subsets regression and stepwise regression.

"Best" Subsets Algorithms

1. Time-saving algorithms require the calculation of only a _____ of all possible regression models.

2. For instance, the algorithms search for the five best subsets of X variables with the smallest C_p values using much less computational effort than when all possible subsets are evaluated. These algorithms are called _____.

3. When the pool of potential X variables is very large, say greater than 30 or 40, even the "best" subset algorithms may require _____.

4. As previously emphasized, our objective at this stage is not to identify _____ _____; we hope to identify a small set of _____ for further study.

5. Example The Surgical Unit Example (eight predictors), we know there are $2^8 = 256$ possible models.

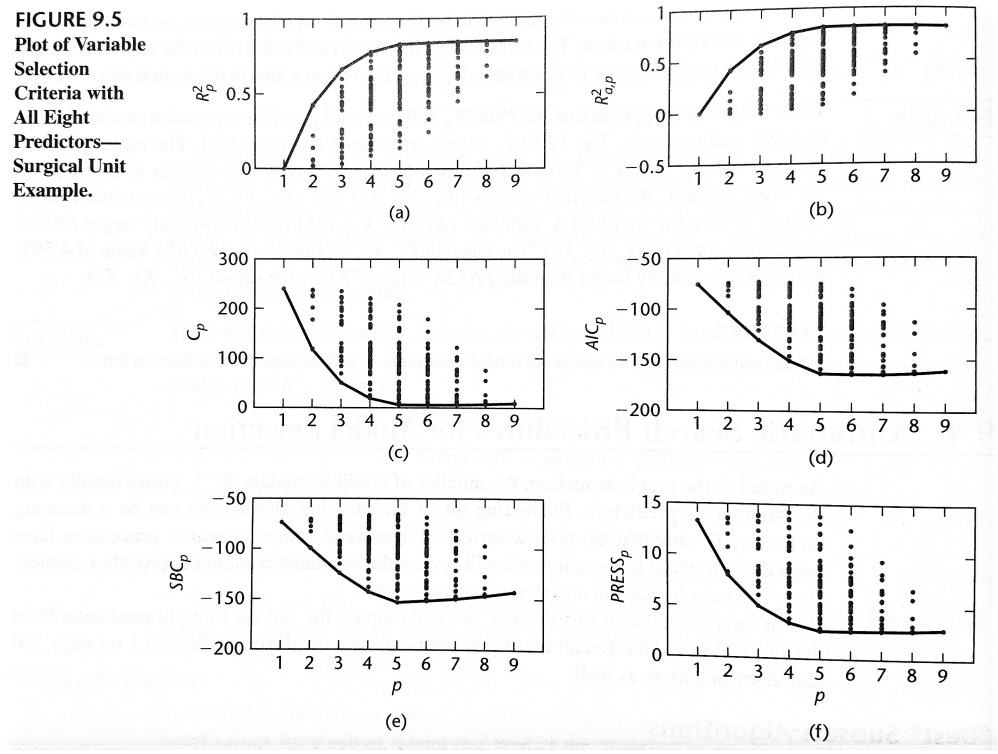


FIGURE 9.5 Plot of Variable Selection Criteria with All Eight Predictors—Surgical Unit Example.

(a) (Figure 9.5) Plots of the six model selection criteria. The best values of each criterion for each p have been connected with _____ lines.

(b) (Table 9.3) The overall _____ criterion values have been underlined in each column of the table.

TABLE 9.3 Best Variable-Selection Criterion Values—Surgical Unit Example.

p	(1) SSE_p	(2) R_p^2	(3) $R_{a,p}^2$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	−75.703	−73.714	13.296
2	7.332	0.428	0.417	117.409	−103.827	−99.849	8.025
3	4.312	0.663	0.650	50.472	−130.483	−124.516	5.065
4	2.843	0.778	0.765	18.914	−150.985	−143.029	3.469
5	2.179	0.830	0.816	5.751	−163.351	−153.406	2.738
6	2.082	0.837	0.821	5.541	−163.805	−151.871	2.739
7	2.005	0.843	0.823	5.787	−163.834	−149.911	2.772
8	1.972	0.846	0.823	7.029	−162.736	−146.824	2.809
9	1.971	0.846	0.819	9.000	−160.771	−142.870	2.931

(c) For example

i. a 7-or 8-parameter model is identified as best by the $R_{a,p}^2$ criterion (both have _____)

ii. a 6-parameter model is identified by the C_p criterion (_____),

iii. a 7-parameter model is identified by the AIC_p criterion (_____).

iv. Both the SBC_p and $PRESS_p$ criteria point to 5-parameter models (_____ and _____).

(d) (Figure 9.6) MINITAB output for the "best" subsets algorithm. We specified that the _____ be identified for each number of variables in the regression model.

FIGURE 9.6 MINITAB Output for "Best" Two Subsets for Each Subset Size—Surgical Unit Example.

Response is lnSurviv

Vars	R-Sq	R-Sq(adj)	C-p	S	Blood	Prog	Enz	Liver	Age	Gender	Hist	Alcohol
1	42.8	41.7	117.4	0.37549			X					
1	42.2	41.0	119.2	0.37746				X				
2	66.3	65.0	50.5	0.29079		X	X					
2	59.9	58.4	69.1	0.31715			X	X				
3	77.8	76.5	18.9	0.23845		X	X					X
3	75.7	74.3	25.0	0.24934	X	X	X					
4	83.0	81.6	5.8	0.21087	X	X	X					X
4	81.4	79.9	10.3	0.22023		X	X	X				X
5	83.7	82.1	5.5	0.20827	X	X	X		X			X
5	83.6	81.9	6.0	0.20931	X	X	X			X		X
6	84.3	82.3	5.8	0.20655	X	X	X			X	X	X
6	83.9	81.9	7.0	0.20934	X	X	X				X	X X
7	84.6	82.3	7.0	0.20705	X	X	X			X	X	X X
7	84.4	82.0	7.7	0.20867	X	X	X	X	X	X		X
8	84.6	81.9	9.0	0.20927	X	X	X	X	X	X	X	X

(e) The MINITAB algolithm uses the _____ criterion, but also shows for each of the "best" subsets the $R^2_{a,p}, C_p$, and $\sqrt{MSE_p}$ (labeled S) values. The rightmost columns of the tabulation show the _____ in the subset.

(f) According to the $R^2_{a,p}$ criterion, the 7-parameter model based on all predictors except _____ (X_4) and _____ (history of moderate alcohol use X_7), or the 8-parameter model based on all predictors except _____ (X_4) are best.

(g) The $R^2_{a,p}$ criterion value for both of these models is _____.

6. The _____ leads to the identification of a small number of subsets that are "good" according to a specified criterion.

7. Consequently, one may wish at times to consider _____ in evaluating possible subsets of X variables.

8. Once the investigator has identified a few "good" subsets for intensive examination, a final choice of the model variables must be made. This choice is aided by _____ (and other _____ to be covered in Chapter 10) and by the investigator's _____ of the subject under study, and is finally confirmed through _____ studies.

Stepwise Regression Methods

1. When the pool of potential X variables contains 30 to 40 or even more variables, use of a "best" subsets algorithm may not be _____.

2. An _____ search procedure that develops the "best" subset of X variables _____ may then be helpful. The _____ procedure is probably the most widely used of the automatic search methods.

3. Essentially, the forward stepwise search method develops _____ _____, at each step _____ or _____ an X variable. The criterion for adding or deleting an X variable can be stated equivalently in terms of _____, coefficient of partial correlation, _____ statistic, or _____ statistic.

4. An essential difference between stepwise procedures and the "best" subsets algorithm is that stepwise search procedures end with the identification of a _____ regression model as "best." With the "best" subsets algorithm, _____ regression models can be identified as "good" for final consideration.

Forward Stepwise Regression

We shall describe the forward stepwise regression search algorithm in terms of the _____ (2.17) and their associated _____ for the usual tests of regression parameters.

1. The stepwise regression routine first fits a _____ model for each of the $p-1$ potential X variables. For each SLR model, the t^* statistic for testing whether or not the slope is zero is obtained:

_____

(a) The X with the _____ value is the candidate for first _____. If this t^* value exceeds a _____, or if the corresponding P-value is less than a predetermined α, the X variable is _____.

(b) Otherwise, the program terminates with _____ considered sufficiently helpful to enter the regression model.

2. Assume X_7 is the variable entered at step 1. The stepwise regression routine now fits all regression models with _____, where X_7 is one of the pair.

(a) For each such regression model, the _____ corresponding to the newly added predictor X_k is obtained.

(b) This is the statistic for testing whether or not _____ when _____ are the variables in the model.

(c) The X variable with the _____ value-or equivalently, the _____is the candidate for addition at the second stage.

(d) If this t^* value exceeds a predetermined level (i.e., the P-value falls below a predetermined level), the second X variable is _____. Otherwise, the program terminates.

3. Suppose X_3 is added at the second stage. Now the stepwise regression routine examines whether any of the other X variables _____ should be _____.

(a) There is at this stage only one other X variable in the model, X_7 , so that only one t^* test statistic is obtained:

$$t_7^* = \_\_\_\_\_$$

(b) At later stages, there would be a number of these t^* statistics, one for each of the variables in the model _____.

(c) The variable for which this _____ (or equivalently the variable for which the P-value is largest) is the candidate for _____.

(d) If this t^* value falls below-or the P-value exceeds-a predetermined limit, the variable is dropped from the model; otherwise, it is _____.

4. Suppose X_7 is retained so that both X_3 and X_7 are now in the model.

   (a) The stepwise regression routine now examines which X variable is the next candidate for _____.

   (b) Then examines whether any of the variables _____ should now be dropped.

   (c) And so on until no further X variables can either be added or deleted, at which point the search _____.

5. Note that the stepwise regression algorithm allows an X variable, brought into the model at an _____ stage, to be dropped subsequently if it is _____ in conjunction with variables added at later stages.

Example

(Figure 9.7) MINITAB computer printout for the forward stepwise regression procedure for The Surgical Unit Example. The maximum acceptable a limit for _____ a variable is 0.10 and the minimum acceptable a limit for _____ a variable is 0.15.

FIGURE 9.7 MINITAB Forward Stepwise Regression Output—Surgical Unit Example.

```
Alpha-to-Enter: 0.1  Alpha-to-Remove: 0.15

Response is lnSurviv on  8 predictors, with N = 54

    Step          1       2       3       4
Constant      5.264   4.351   4.291   3.852

Enzyme       0.0151  0.0154  0.0145  0.0155
T-Value        6.23    8.19    9.33   11.07
P-Value       0.000   0.000   0.000   0.000

ProgInde             0.0141  0.0149  0.0142
T-Value                5.98    7.68    8.20
P-Value               0.000   0.000   0.000

Histheav                     0.429   0.353
T-Value                       5.08    4.57
P-Value                      0.000   0.000

Bloodclo                             0.073
T-Value                               3.86
P-Value                              0.000

S            0.375   0.291   0.238   0.211
R-Sq         42.76   66.33   77.80   82.99
R-Sq(adj)    41.66   65.01   76.47   81.60
C-p          117.4    50.5    18.9     5.8
```

1. At the start of the stepwise search, _____ is in the model so that the model to be fitted is $Y_i = \beta_0 + \epsilon_i$;.

   (a) (Step 1), the _____ statistics and corresponding P-values are calculated for each potential X variable, and the predictor having the _____ (_____) is chosen to enter the equation.

   (b) Enzyme (X_3) had the largest test statistic:

   $$t_3^* = \frac{b_3}{s\{b_3\}} = \frac{0.015124}{0.002427} = \_\_\_\_\_.$$

   (c) The P-value for this test statistic is _____, which falls below the maximum acceptable α-to-enter value of 0.10; hence Enzyme (X_3) is added to the model.

   (d) The current regression model contains Enzyme (X_3), "Step 1": the regression coefficient for Enzyme (0.0151).

   (e) At the bottom of column 1, a number of variables-selection criteria, including $R_1^2(42.76)$, $R_{a,1}^2(41.66)$, and $C_1(117.4)$ are also provided.

2. Next, all regression models containing X_3 and _____ variable are fitted, and the t^* statistics calculated:

   $t_k^* =$ _____, since _____, _____

   Progindex (X_2) has the highest t^* value, and its P-value (0.000) falls below 0.10, so that X_2 now enters the model.

3. Enzyme and Progindex (X_3 and X_2) are now in the model. At this point, a test whether _____ should be dropped is undertaken, but because the _____ (0.000) corresponding to X_3 is not above 0.15, this variable is _____.

4. Next, all regression models containing X_2, X_3, and one of the remaining potential X variables are fitted. The appropriate t^* statistics:

   $$t_k^* = \_\_\_\_\_$$

   The predictor labeled Histheavy (X_8) had the largest t^* value, (P-value = 0.000) and was next added to the model. X_2, X_3 , and X_8 are now in the model.

5. Next, a test is undertaken to determine whether _____. Since both of the corresponding P-values are less than 0.15, neither predictor is dropped from the model.

6. (Step 4) Bloodclot (X_1) is added, and no terms previously included were dropped. The right-most column of Figure 9.7 summarizes the addition of variable X_1 into the model containing variables X_2, X_3, and X_8.

7. Next, a test is undertaken to determine whether either _____ should be dropped. Since all P-values are less than 0.15 (all are 0.0(0), all variables are retained.

8. Finally, the stepwise regression routine considers adding one of X_4 , X_5 , X_6 , or X_7 to the model containing X_1, X_2, X_3, and X_8. In each case, the P-values are greater than 0.10 (not shown); therefore, no additional variables can be added to the model and the search process is terminated.

9. Thus, the stepwise search algorithm identifies _____ as the "best" subset of X variables. This model also happens to be the model identified by both the _____ and _____ criteria in our previous analyses based on an assessment of "best" subset selection.

Other Stepwise Procedures

1. *Forward Selection*. The forward selection search procedure is a simplified version of forward stepwise regression, _____ whether a variable once entered into the model should be _____.

2. *Backward Elimination*. The backward elimination search procedure is the _____ _____ selection.

   (a) It begins with the model containing _____ potential X variables and identifies the one with the largest P-value.

   (b) If the maximum P-value is greater than a predetermined limit, that X variable is dropped.

(c) The model with the remaining $(P-2)$ X variables is then fitted, and the next candidate for dropping is identified.

(d) This process continues until no further X variables can be dropped.

9.5 Some Final Comments on Automatic Model Selection Procedures*

9.6 Model Validation

1. The final step in the model-building process is the _____ of the selected regression models.

2. Model validation usually involves checking a _____ against _____. Three basic ways of validating a regression model are:

   (a) Collection of _____ to check the model and its predictive ability.

   (b) _____ of results with theoretical expectations, earlier empirical results, and simulation results.

   (c) Use of a _____ to check the model and its _____.

3. What is difference between: training set, testing set and hold-out set: (The training set is for _____)

   (a) A observed data set (100%): e.g, training set (75%), testing set (25%).

   (b) A observed data set (100%): k-fold cross validation: e.g, $k = 4$ (25%, 25%, 25%, 25%), in turns "testing set (25%), training set (75%)" 4 times.

   (c) A observed data set (100%): hold-out set (20%), Not hold-out set (80% for 4-fold CV)

Collection of New Data to Check Model

1. The _____ means of model validation is through the _____. The purpose of collecting new data is to be able to examine whether the regression model developed from the earlier data is still _____. If

so, one has assurance about the _____ of the model to data beyond tho,se on which the model is based.

Methods of Checking Validity. A means of measuring the _____ of the selected regression model is to use this model to predict each case in the new data set and then to calculate the mean of the squared prediction errors, to be denoted by $MSPR$, which stands for mean squared prediction error:

$$MSPR = \_\_\_\_\_$$

where:

- Y_i is the value of the response variable in the ith _____.
- $\hat{Y}_i$ is the _____ for the ith validation case based on the model-building dataset.
- n^* is the number of cases in the validation data set.

2. If the mean squared prediction error $MSPR$ is fairly close to _____ based on the regression fit to the _____, then the error mean square MSE for the selected regression model is _____ and gives an appropriate indication of the predictive ability of the model.

3. If the mean squared prediction error is _____, one should rely on the mean squared prediction error as an indicator of how well the selected regression model will predict in the future.

😍 TA Class

- **Problems**: 9.6, 9.11, 9.18, 9.21
- **Exercises**: none
- **Projects**: none