Chapter 15 Variable Selection

In constructing a model for Y, we want to use as few predictor variables as possible that adequately explain the relationship between Y and $X_1, \cdots, X_m$.

<u>Reasons</u>: (1) Reduce the effect of collinearity.

(2) Concentrate on the subset of variables most important to the response.

(3) A simple model will be easier to explain and validate on new data.

(4) Reduce study cost.

(5) Keep bias errors small.

(6) Keep the variance of the predictions $\left(\frac{\sum Var(\hat{Y}_i)}{n} = \frac{p\sigma^2}{n}\right)$ small.

Predictor variables under consideration: $X_1, X_2, \cdots, X_m$.

Suppose that we always include an intercept in a model. Then

all possible models $= \underbrace{2 \times 2 \times \cdots \times 2}_{m} = 2^m = \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{m}$.

Ex. 15.1: y = son's height, X_1 = father's height, X_2 = mother's height.

4 possible models are: (1) $y = \beta_0 + \varepsilon$

(2) $y = \beta_0 + \beta_1 X_1 + \varepsilon$

(3) $y = \beta_0 + \beta_2 X_2 + \varepsilon$

(4) $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

If we also consider X_3 = father's weight, X_4 = mother's weight,

all possible models $= 2^4 = 16$.

How to select the best model?

— Generally use knowledge of subject under study to help decide which variables should be included.

— Statistically, the selection is based on R^2, or adjusted R^2, or s^2, or some other important statistics AIC, AIC_c and BIC.

<u>15.1. Stepwise Regression</u>

Principle: Consider adding/deleting variables to/from the model one at a time.

(I) Testing-Based Procedures

— The procedures are based on hypothesis tests.

• Forward Selection (FS)

(1) Start from the empty model: $Y = \beta_0 + \varepsilon$.

(2) Add a new variable if it (a) increases R^2 more than any other variable, or equivalently

(b) has the largest $|t_{obs}|$ when added of any other variable, or equivalently

(c) has the smallest p-value when added of any other variable since p-value $= 2 \times P(t_\nu > |t_{obs}|)$.

(3) Continue with (2) until a stopping rule is met.

Rule 1: Stop with k^* predictors, where k^* is predetermined.

Rule 2: Stop if collinearity becomes a severe problem. (tolerance check — built in all good programs.)

Rule 3: Stop when next new variable is no longer significant at specified level α_F^*.

- <u>Backward Elimination (BE)</u>

(1) Start with all variables in the model, i.e., the full model.

(2) Delete a variable if it is least significant of remaining variables, equivalently it has the largest p-value for the t test.

(3) Continue with (2) until stopping rule is met.

Rule 1: Stop when model has k^* predictors, where k^* is predetermined.

Rule 2: Stop when next variable to be deleted is significant at specified level α_B^*.

- <u>Stepwise Selection (SS)</u>

At each step, consider adding and deleting a variable using FS and BE.

(1) Start from the empty model: $Y = \beta_0 + \varepsilon$.

(2) Add a variable which is most significant at a specified level α_F^*.

(3) Delete a variable which is least significant at a specified level α_B^* (must be $\alpha_B^* \geq \alpha_F^*$).

(4) Continue with (2) and (3) until no change.

Note: Testing-based stepwise variable selection procedures are not encouraged since their overall type I error rates may be very high.

(II) <u>Criterion-Based Procedures</u>

The general form of most information criteria is

$$IC(k) = -2\ln(\text{maximum likelihood}) + kr,$$

where r is the number of parameters involved in the model and k is a penalty coefficient that is chosen in advance and used for all models. $IC(k)$ not only rewards goodness of fit by $-2\ln(\text{maximum likehood})$, but also penalizes the addition of extra predictor variables to prevent overfitting. For a given k, the smaller $IC(k)$ is, the better the model is. The three most common information criteria are:

1. Akaike's Information Criterion (AIC):

$$AIC = IC(k) = -2\ln(\text{maximum likelihood}) + 2r.$$

2. Corrected AIC:

$$AIC_c = IC\left(\frac{2n}{n-r-1}\right) = -2\ln(\text{maximum likelihood}) + \frac{2r(r+1)}{n-r-1}.$$

3. Bayesian Information Criterion (BIC):

$$BIC = IC(\ln(n)) = -2\ln(\text{maximum likelihood}) + r\ln(n).$$

Note: A larger k means that the penalty for additional parameters is more severe, for example, when $\ln(n) > 2$ (equivalently, $n > e^2 \approx 7.39$), the penalty for additional parameters in BIC is more severe than that in AIC.

- Forward Selection (FS)

(1) Select a k for the information criterion.

(2) Start with the empty model, $Y = \beta_0 + \varepsilon$, and compute $IC(k)$ on this model.

(3) Add a new variable if the addition reduces $IC(k)$ most.

(4) Continue with (3) until any further addition will not reduce $IC(k)$.

(5) Report the model with the the smallest $IC(k)$ as the final model.

- Backward Elimination (BE)

(1) Select a k for the information criterion.

(2) Start with the full model, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$, and compute $IC(k)$ on this model.

(3) Eliminate a variable if the elimination reduces IC(k) most.

(4) Continue with (3) until any further elimination will not reduce IC(k).

(5) Report the model with the smallest IC(k) as the final model.

- Stepwise Selection (SS)

At each step, consider adding and deleting a variable using FS and BE.

(1) Select a k for the information criterion.

(2) Start for the empty model, $Y = \beta_0 + \varepsilon$, and compute IC(k) on this model.

(3) Add a variable if the addition reduces IC(k) most.

(4) Elimiate a variable if the elimination reduces IC(k) most.

(5) Continue with (3) and (4) until no further reduction in IC(k).

(6) Report the model with the smallest IC(k) as the final model.

Notes: (i) Advantage of the stepwise procedures: Easy to use, implement and inexpensive to do calculations.

(ii) Problems of the stepwise procedures:

— They do not necessarily lead to the optimal model.

— Ordering of predictors is not necessarily indicative of importance in describing relationship.

— Can lead to overfitting the data.

5.2. <u>All Possible Regressions</u>

As we discussed at the beginning of this chapter, if m predictor variables are under consideration, there are 2^m possible models. If we consider all those possible models and find the best one, the approach is called <u>all possible regressions</u>.

Advantage of all possible regressions: Can get the best model based on a criterion.

Problem of the method: All possible regressions are not feasible for a large m, for example, if $m = 20$, more than 1 million models need to be fitted.

Ex. 15.2: Fit an appropriate model for the steam plant data in Example 5.3 by the forward selection procedure, the backware elimination procedure, and the stepwise selection procedure based on BIC.