Regression Variable Selection

The model-building for regression is an interactive process. Roughly speaking, it consists of four phases

1. Data collection
2. Reduction of predictor variables
3. Model refinement and selection
4. Model validation

For different purposes of the study we often have different models. E.g. if the purpose is only **prediction**, generally we want to include a subset of the variables in the model. But if the purpose is **explanation**, e.g. trying to build a descriptive model (often used for causal inference in some fields), then probably the more variables we have the better. Due to the correlations among predictor variables these two purposes, prediction and variable selection, often can not be satisfied simultaneously. Generally the inclusion of multiple correlated variables in the model will make the individual regression coefficients to be estimated unstably, i.e. with big variance. If we only use a small subset of variables, then the models are likely to be biased. Often we would like to achieve a balance between the variance and the bias. We can certainly say that

"all models are wrong, but some are useful".

As the examples in our last chapter illustrate, the including of highly correlated variables in the model often causes the correlated predictors to be "not significant" marginally. So the commonly used practice of selecting "important predictors" based on the marginal t-statistics

$$t_k^* = \frac{b_k}{s(b_k)}$$

in a big regression model is not appropriate. We will introduce different criteria for variable selection for different purposes of the studies.

(The textbook has a very good overall description about the model-building process about different studies.)

Model Selection Criteria

Suppose we have a total P-1 predictors, and we want to select a subset of them (p-1 variables) to be included in our final model. And we assume the sample size n is big enough.

Question:

1. How many different models ?
2. How many different linear models?
3. How to determine the "best" model?

R_p^2 or SSE_p Criterion

$$R_p^2 = 1 - \frac{SSE_p}{SSTO} \tag{1}$$

is the coefficient of multiple determination for the regression model with p predictors. It is the proportion of variation in response y explained by the p predictors.

We are not intending to identify the subsets which maximize R_p^2 since it is always increasing with more predictors. The intent is to find the point where adding more predictors is not worthwhile because it leads to a very small increase in R_p^2.

MSE_p or $R^2_{a,p}$ Criterion

R^2_p does not take into account the number of parameters in the model. The adjusted coefficient of multiple determination

$$R^2_{a,p} = 1 - \frac{SSE_p/(n-p)}{SSTO/(n-1)} = 1 - \frac{MSE_p}{SSTO/(n-1)} \tag{2}$$

is suggested as an alternative criterion.

Mallows' C_p Criterion

This criterion is concerned with the total mean squared error of the n fitted values for each subset of predictors. The mean squared error concept involves the total error in each fitted values

$$\hat{Y}_i - \mu_i,$$

where μ_i is the true mean response when the levels of the predictor variables are those for the ith case.

It can be decomposed into a bias and a random error component

$$\hat{Y}_i - \mu_i = \hat{Y}_i - E(\hat{Y}_i) + E(\hat{Y}_i) - \mu_i,$$

where

1. $E(\hat{Y}_i) - \mu_i$ is the bias component of the ith fitted values $\hat{Y}_i$, also called the model error component. The intuition is that if the fitted model is not correct, $E(\hat{Y}_i)$ will differ from the true mean response μ_i and the difference represents bias of the fitted model.

2. $\hat{Y}_i - E(\hat{Y}_i)$ is the random error component for $\hat{Y}_i$. It represents the deviation of the fitted values $\hat{Y}_i$ for the ith sample from the expected values when the ith fitted values is obtained by fitting the same regression model to all samples.

The mean squared error of $\hat{Y}_i$ is defined as the expected value of the square of the total error as following

$$(\hat{Y}_i - \mu_i)^2 = \left\{[\hat{Y}_i - E(\hat{Y}_i)] + [E(\hat{Y}_i) - \mu_i]\right\}^2$$

The criterion is defined as

$$\Gamma_p = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\hat{Y}_i - \mu_i)^2. \tag{3}$$

Notice that

$$(\hat{Y}_i - \mu_i)^2 = (\hat{Y}_i - Y_i)^2 + (Y_i - \mu_i)^2 - 2(\hat{Y}_i - Y_i)(Y_i - \mu_i),$$

we have

$$E\{(Y_i - \mu_i)^2\} = \sigma^2,$$

$$\begin{aligned} E\{(Y_i - \hat{Y}_i)(Y_i - \mu_i)\} &= E\{Y'(I - H)(Y - \mu)\} \\ &= tr\{E[(Y - \mu)Y'(I - H)]\} \\ &= (n - p)\sigma^2, \end{aligned}$$

so we can use

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p) \tag{4}$$

to estimate Γ_p, where σ^2 is estimated from the full model.

AIC_p and SBC_p/BIC_p Criteria

$$AIC_p = n\log(SSE_p) - n\log(n) + 2p \tag{5}$$

$$BICp = n\log(SSE_p) - n\log(n) + \log(n)p \tag{6}$$

$PRESS_p$ Criterion

The PRESS prediction error for the ith case is

$$Y_i - \hat{Y}_{i(i)},$$

where $\hat{Y}_{i(i)}$ is obtained by using the model built from all the samples except the ith case. The $PRESS_p$ (prediction sum of squares) is

$$PRESS_p = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2, \tag{7}$$

where h_{ii} is the iith element of hat matrix H.

Cross Validation (CV)

- $PRESS_p$ is also called leave-one-out-cross-validation (LOOCV).
- Instead of LOOCV, we can consider general K-fold CV (KFCV), where all samples are randomly divided into K groups. Each time one of the K-fold is left out (as a testing set), all the other $K-1$ fold (training set) are used to build a linear regression model and used to predict the left-out fold. Final prediction error is the average over the K-fold.
- Randomized CV

Sequential Search Procedure

Forward stepwise regression is developed to reduce the computational efforts as compared with the all-possible-regression procedure. It is a sequential procedure, at each step adding or deleting an X variable. The criterion for adding or deleting an X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t statistic, F statistic. It works in the following way.

Forward Stepwise Selection Procedure:

1. Fit a simple linear regression model for each of the $P-1$ potential X variables, and choose the one with the best value of previous criteria, say X_{k_1} is added in the model.

2. With X_{k_1} in the model, we add in another one by maximizing the criterion. X_{k_2} is added when its value is the best and better than some (addin) threshold.

3. Drop X_{k_1} from the model if its value is less than some (dropout) threshold.

4. Repeat previous addin and dropout steps until no variable can be added or dropped.

Forward selection search procedure simplified the forward stepwise regression, omitting the test whether a variable once entered into the model should be dropped.

Backward elimination search procedure begins with the model containing all potential X variables and considers dropping variables sequentially. A stepwise modification can be adapted that enables variables eliminated earlier to be added later. This modified procedure is called backward stepwise regression procedure.

Model Validation

- Collect new data
- Compare with known results
- Data spliting