6.1 Subset Selection

Zongyi Liu

2023-05-25

6.1 Subset Selection

6.1.1 Best Subset Selection

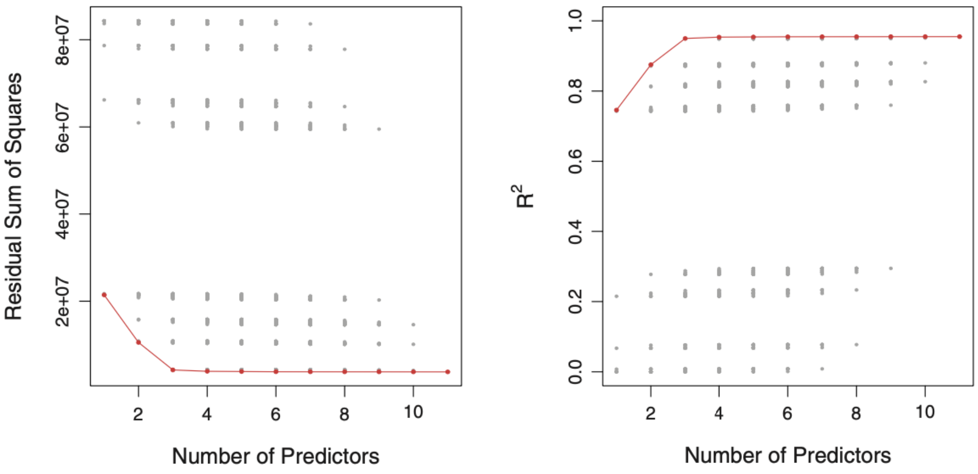
To perform best subset selection, we fit a separate least squares regression for each possible combination of p predictors. We can use the algorithm below to perform:

Algorithm 6.1 Best Subset Selection

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, .., p$: first to find all $\binom{p}{k}$ models that contain exactly k predictors; Pick the best among these $\binom{p}{k}$ models, and call it M_k. Here the best is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from $M_0, ..., M_p$ using cross validated prediction error, C_p (AIC), BIC or adjusted R^2.

In order to select a single best model, we must choose among $p + 1$ options. We should notice that the RSS of these $p + 1$ models decrease monotonically, and the R^2 increases monotonically.

We should also notice that a low RSS or a high R^2 indicates a model with a low training error, whereas we wish to choose a model with a low test error. Therefore, in step 3, we need to use CV prediction error, C_p, BIC, or adjusted R^2 in order to select among M_0, M_1,. . . , M_p.



For each possible model containing a subset of the ten predictors in the `Credit` data set, the RSS and R^2 are displayed as above.

6.1.2 Stepwise Selection

The best subset selection cannot be applied with a very large p, which might lead to statistical problems.

Forward Stepwise Selection

Algorithm 6.2 Forward Stepwise Selection

1. Let M_0 denote the null model, which contains no predictors
2. For $k = 0, 1, 2, .., p - k$:

- consider all $p - k$ models that augment the predictors in M_k with one additional predictors
- Choose the best among these $p - k$ models, and call it M_{k+1}. Here the best is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from $M_0, ..., M_p$ using cross validated prediction error, C_p (AIC), BIC or adjusted R^2.

Backward Stepwise Selection

Algorithm 6.3 Backward Stepwise Selection

1. Let M_p denote the null model, which contains all p predictors.
2. For $k = p, p - 1, ..., 1$

- Consider all k models that contain all but one of the predictors in M_k, for a total of $k - 1$ predictors
- Choose the best among these k models, and call it M_{k-1}. Here the best is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from $M_0, ..., M_p$ using cross validated prediction error, C_p (AIC), BIC or adjusted R^2.

Like forward stepwise selection, the backward selection approach searches through only $1+p(p+1)/2$ models, and so can be applied in settings where p is too large to apply best subset selection. It also requires that the number of samples n is larger than the number of variable p, whereas the forward stepwise selection is viable even when $n < p$

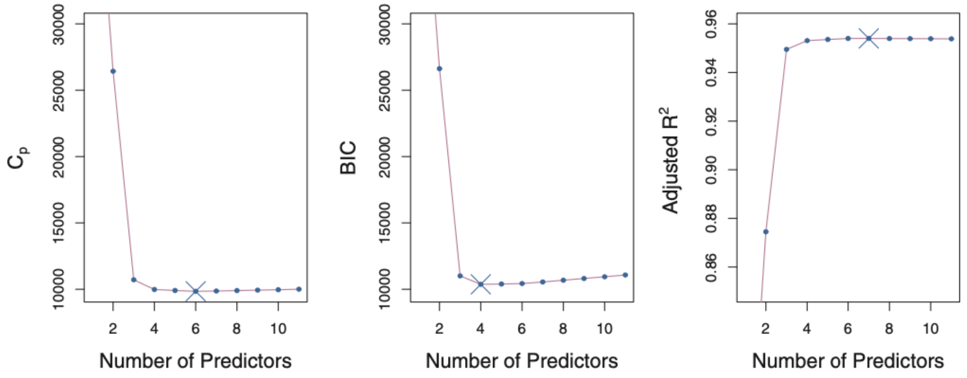
6.1.3 Choosing the Optimal Model

We need to consider which model is better when trying to fit the data. To reach this goal, we need to estimate the test error. There are two common approaches:

1. Indirectly estimate the test error by making an adjustment to the training error to account for the bias due to overfitting
2. Directly estimate the test error using either a VS approach or a CV approach

C_p, AIC, BIC, and Adjusted R^2

We show in Chapter 2 that the training set MSE is generally an underestimate of the test MSE. (Recall that MSE = RSS/n.) This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS (but not the test RSS) is as small as possible.



In this plot, C_p, BIC, and adjusted R2 are shown for the best models of each size for the Credit data set (the lower frontier in Figure 6.1). C_p and BIC are estimates of test MSE. In the middle plot we see that the BIC estimate of test error shows an increase after four variables are selected. The other two plots are rather flat after four variables are included.

For fitting a least squares model containing d predictors, the C_p estimate of MSE is computed as

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

The AIC criterion is defined for a large class of models fit by maximum likelihood. In the case of the model above with Gaussian errors, maximum likelihood and least squares are the same thing. In this case AIC is given by

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{sigma}^2)$$

BIC is derived from a Bayesian point of view, but ends up looking similar to C_p (and AIC) as well. For the least squares model with d predictors, the BIC is, up to irrelevant constants, given by

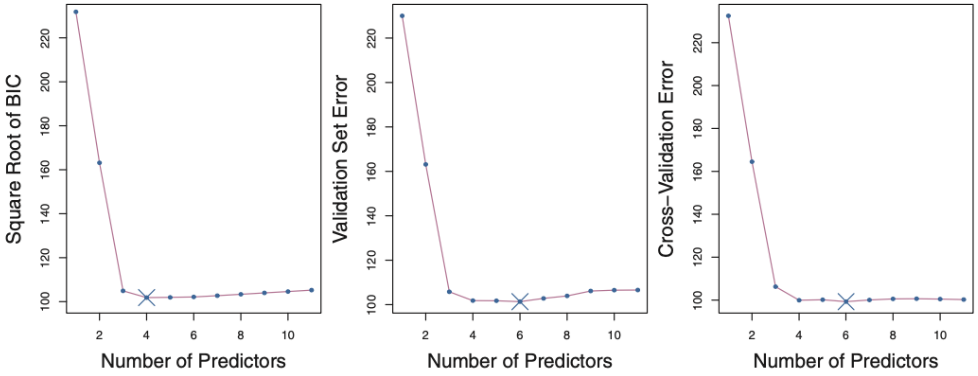
$$BIC = \frac{n}{\hat{\sigma}^2}(RSS + log(n)d\hat{\sigma}^2)$$

Generally we select the model that has the lowest BIC value. Notice that BIC replaces the $2\hat{\sigma}^2$ used by C_p with a $log(n)d\hat{\sigma}^2$ term, where n is the number of observations. Since log $n > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than C_p.

The adjusted R^2 statistic is another popular approach for selecting among a set of models that contain different numbers of variables. Recall from Chapter 3 that the usual R^2 is defined as $1 - RSS/TSS$.

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

In this plot below, we can see as a function of d, the BIC, validation set errors, and cross-validation errors on the Credit data, for the best d-variable model. The validation errors were calculated by randomly selecting three-quarters of the observations as the training set, and the remainder as the validation set. The cross-validation errors were computed using k = 10 folds. In this case, the validation and cross-validation methods both result in a six-variable model. However, all three approaches suggest that the four-, five-, and six-variable models are roughly equivalent in terms of their test errors.


Square Root of BIC
Number of Predictors
Validation Set Error
Number of Predictors
Cross−Validation Error
Number of Predictors