Model Selection

Model Selection

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response.
- Including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables—that is, by setting the corresponding coefficient estimates to zero—we can obtain a model that is more easily interpreted.
- Now least squares is extremely unlikely to yield any coefficient estimates that are exactly zero.

Subset Selection

- some approaches for automatically performing feature selection or variable selection—that is, for excluding irrelevant variables from a multiple regression model are presented.
- The Subset Selection is one of the way to achieve this.

Subset Selection

- Subset Selection. This approach involves identifying a subset of the p predictors that we believe to be related to the response.
- We then fit a model using least squares on the reduced set of variables.

Subset Selection- Best Subset Selection

- To perform best subset selection, we fit a separate least squares regression for each possible combination of the p predictors.
- That is, we fit all 2^p models selection that contain exactly one predictor, all models that contain exactly two predictors, and so forth.
- Then look at all of the resulting models, with the goal of identifying the one that is best.

Subset Selection- Best Subset Selection

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

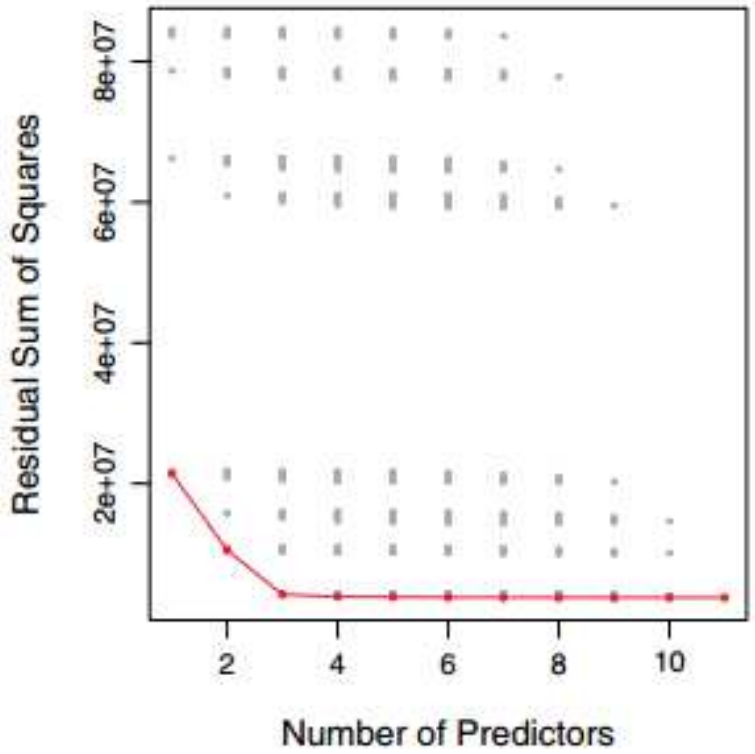
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Subset Selection- Best Subset Selection

- Step 2 identifies the best model (on the training data) for each subset size, in order to reduce the problem from one of 2p possible models to one of p + 1 possible models.



Subset Selection- Best Subset Selection

- To select a single best model, we must simply choose among these p + 1 options.
- This must be performed with care, as the RSS of these p + 1 models decreases monotonically, as the number of features included in the models increases.
- If we use these statistics to select the best model, then we will always end up with a model involving all of the variables.

Subset Selection- Best Subset Selection

- The problem is that a low RSS indicates a model with a low training error, whereas we wish to choose a model that has a low test error.
- Therefore, in Step 3, we use cross-validated prediction error, C_p, BIC, or adjusted R^2 in order to select among $M_0, M_1, \ldots, M_p$.

Subset Selection- Best Subset Selection

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

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

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right).$$

Limitations- Best Subset Selection

1. Suffers from computational limitations.
   - The number of possible models that must be considered grows rapidly as p increases.
   - In general, there are 2^p models that involve subsets of p predictors.
   - If p = 10, then there are approximately 1,000 possible models to be considered.
   - best subset selection becomes computationally infeasible for values of p greater than around 40, even with extremely fast modern computers.

Limitations- Best Subset Selection

2. They also only work for least squares linear regression.

Stepwise Selection

- a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure.
- In each step, a variable is considered for addition to or subtraction from the set of variables based on some prespecified criterion.
- The two approaches are: Forward Stepwise Selection and Backward Stepwise Selection.

Forward Stepwise Selection

- Forward stepwise selection is a computationally efficient alternative to best subset selection.
- Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model.
- In particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model.

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward stepwise selection

- Forward stepwise selection involves fitting one null model, along with p – k models in the kth iteration, for k = 0, . . . , p – 1. This amounts to a total of 1+ p*(p+1) models.
- In Step 2(b) of Algorithm 6.2, we must identify the best model from among those p–k that augment Mk with one additional predictor.This is done by simply choosing the model with the lowest RSS.
- Step 3, we must identify the best model among a set of models with different numbers of variables.

Forward stepwise selection

- Forward stepwise is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
- For instance, suppose that in a given data set with p = 3 predictors, the best possible one-variable model contains X1, and the best possible two-variable model instead contains X2 and X3. Then forward stepwise selection will fail to select the best possible two-variable model, because M_1 will contain X1, so M_2 must also contain X1 together with one additional variable.

Forward stepwise selection

- Forward stepwise selection can be applied even in the high-dimensional setting where $n < p$; however, in this case, it is possible to construct submodels $M_0, \ldots, M_{n-1}$ only, since each submodel is fit using least squares, which will not yield a unique solution if $p \geq n$

Backward stepwise selection

- It begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.
- The backward selection approach searches through only $1+p(p+1)/2$ models.
- Backward selection requires that the number of samples n is larger than the number of variables p (so that the full model can be fit). In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Backward stepwise selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.