# Predictive Analytics
# Regression and Classification
## Lecture 2 : Part 5

**Sourish Das**

Chennai Mathematical Institute

Aug-Nov, 2020

# Feature Selection (aka. Variable Selection)

- Suppose the feature space has $p$ many features, i.e.,

$$\mathbf{X} = \{X_1, X_2, \cdots, X_p\}$$

  $p$ is very large.

- We would like to drop features which have no impact on $\mathbf{y}$

$$\mathbf{y} = f(X_1, \cdots, X_q)$$

  where $q << p$
- Ex: $p = 2000$ and $q = 15$

# Best Subset Selection

- To perform *best subset selection*, we fit a separate least squares regression best subset for each possible combination of the $p$ predictors.
- That is, we fit all $p$ models that contain exactly one predictor, all ${}^{p}C_2 = p(p-1)/2$ models that contain exactly two predictors, and so forth. We then look at all of the resulting models, with the goal of identifying the one that is best.
- The size of the model space is $2^p - 1$.

# Best Subset Selection

**Algorithm**:

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 $k = 1, 2, \cdots, p$;
   2.1 Fit all ${}^pC_k$ models that contain exactly $k$ predictors.
   2.2 Pick the best among these ${}^pC_k$ 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 $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$, using crossvalidated prediction error, $AIC$, $BIC$, or adjusted $R^2$.

# Best Subset Selection

1. Though the step 2 identifies the best model (on the training data) for each subset size, in order to reduce the problem from one of $2^p$ possible models to one of the $p+1$ possible models.
2. The best subset selection involves fitting of $2^p$ models.
3. When $p = 20$, the best subset selection requires fitting $1,048,576$ models.
4. This means best subset selection is almost not possible, unless it is a toy/small dataset.

# Forward stepwise selection

**Algorithm**:

1. Let $\mathcal{M}_0$ denote the null model , which contains no predictors.
2. For $k = 0, 1, \cdots, p - 1$;
   2.1 Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   2.2 Pick the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having the smallest $RSS$, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$, using crossvalidated prediction error, $AIC$, $BIC$, or adjusted $R^2$.

# Forward stepwise selection

- Unlike best subset selection, which involved fitting $2^p$ models, *forward stepwise selection* involves fitting one null model, along with $p - k$ models in the $k^{th}$ iteration, for $k = 0, \cdots, p - 1$.
- This amounts to a total of $1 + \sum_{k=0}^{p-1}(p - k) = 1 + p(p + 1)/2$ models.
- This is a substantial difference: when $p = 20$, *best subset selection* requires fitting $1, 048, 576$ models, whereas forward stepwise selection requires fitting only 211 models.
- 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 $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_{n-1}$ only, since each submodel is fit using least squares, which will not yield a unique solution if $p \geq n$.

# Backward stepwise selection

**Algorithm**:

1. Let $\mathcal{M}_p$ denote the full model , which contains all predictors.
2. For $k = p, p-1, \cdots, 1$;
   2.1 Consider all $k$ models that contain all but one predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   2.2 Pick the best among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having the smallest $RSS$, or equivalently largest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \mathcal{M}_1, \cdots, \mathcal{M}_p$, using crossvalidated prediction error, $AIC$, $BIC$, or adjusted $R^2$.

# Backward stepwise selection

- 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.
- Also like *forward* stepwise selection, *backward* stepwise selection is not guaranteed to yield the *best model* containing a subset of the $p$ predictors.

# Implementation

- In `R`, the built-in function called `step` in `stats` package, select a model by AIC in a Stepwise Algorithm.
- Several `Python` implementation of step-wise feature selection is also available.

# In the next lecture...

- We will discuss the issues of multicollinearity and more...

Thank You

sourish@cmi.ac.in