Linear Model Selection

Sravani Vadlamani

Agenda

- Linear Model Selection
  - Subset selection
  - Stepwise selection
  - Shrinkage methods
  - Dimension reduction

Linear Regression

- $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots \beta_p X_p + \varepsilon$
- Above linear model is fit using least squares
- Higher order regression models and non-linear models will be discussed in future topics
- Linear models are easy to interpret
- This chapter will focus on improving linear models using alternative fitting procedures
- Why improve linear models
  - Better prediction accuracy
  - Improve model interpretability

Prediction Accuracy

- $Least\ squares\ estimates\ will\ have\ low\ bias\ if\ the\ relation\ between\ the\ response\ and\ predictor\ variables\ is\ truly\ linear.$
- If the number of observations is much larger than the number of variables, the least squares estimates will have low variance.
- If the number of observations is not much larger than the number of variables, there can be lot of variability in least squares estimates resulting in overfitting and poor predictions.
- When there are fewer observations than variables, least squares cannot be used. By constraining or shrinking the estimated coefficients, the variance can be reduced at the cost of a negligible increase in bias. This results in improvement in prediction accuracy on new data.

Model Interpretability

- Some or many of the variables included in a multiple regression model may not be associated with the response variable.
- Including irrelevant variables increases model complexity and reduces model interpretability.
- Feature selection or variable selection is used to exclude irrelevant variables from a regression model

Best Subset Selection

- Fit a separate least squares regression model for each possible combination of the *p* predictors i.e., fit *p* models with exactly one predictor, $\binom{p}{2}$ models with two predictors and so on
- Selecting the best model from the possible 2^p models is not a trivial process and involves the following steps:
  1. Let M_0 denote the null model with no predictors that predicts the sample mean for each observation
  2. For k = 1, 2, …..p:
     1. Fit all $\binom{p}{k}$ models that contain exactly k predictors
     2. Choose the best model M_k that yields the smallest RSS or equivalently largest R^2
  3. Select the best model from $M_0, \ldots\ldots\ldots M_p$ using Cp, Akaike Information Criterion (AIC), BIC or adjusted R^2.

Best Subset Selection

- Step 2 above tries to reduce the problem from one of 2^p models to p+1 models
- The best model must be chosen from the p+1 models and should be done with care because with increase in number of variables, R^2 increases monotonically, and RSS decreases monotonically. Due to this, picking the best model will involve a model with all variables. RSS and R^2 are measures of training error and it is better to select a model that has a low test error.
- Hence, we use C_p, BIC or adjusted R2 to select the best model.

Best Subset Selection

- Best subset selection can also be applied to logistic regression. Instead of RSS, we use deviance (negative two times the maximum likelihood) to assess model performance. The smaller the deviance, the better the fit.
- Limitations
  - Simple concept but suffers from computational limitations
  - As *p* increases, number of models increases
  - Computationally infeasible for values *p* greater than 40
  - Leads to overfitting and high variance of coefficient estimates
  - Techniques like branch-and-bound help eliminate some choices but work only for least squares linear regression

Stepwise Selection

- An attractive alternative to best subset selection that explore a far more restricted set of models
- Includes two approaches
  - Forward stepwise selection
  - Backward stepwise selection

Forward Stepwise Selection

- Begins with a model with no predictors and adds predictors one at a time till all the predictors are included. At each step, the variable that gives the greatest additional improvement to the fit is added to the model
- Algorithm includes the following steps:
  1. Let M_0 denote the null model with no predictors that predicts the sample mean for each observation
  2. For k – 1, 2, …..p-1:
     1. Consider all (p-k) models that augment the predictors in M_k with one additional parameter
     2. Choose the best among (p-k) models that yields the smallest RSS or largest R^2 and call it M_{k+1}
  3. Select the best model from $M_{0,}$ …… M_p using Cp, Akaike Information Criterion (AIC), BIC or adjusted R^2.

Forward Stepwise Selection

- Estimates 1 + p(p+1)/2 models which is a substantial improvement over best subset selection
- Forward stepwise selection may not always find the best possible model out of the 2^p models due to its additive nature. For example, forward stepwise selection could not find the best 2-variable model in a data set where the best 1-variable model utilizes a variable not used by the best 2-variable model
- Can be applied in high dimensional scenarios when n <p, however it can construct only n-1 models as it uses least squares regression which will not yield a unique solution if p>=n

Backward Stepwise Selection

- Begins with a model with all the predictors and iteratively remove the lest useful predictor one at a time.
- Algorithm includes the following steps:
  1. Let M_p denote the full model with all the predictors
  2. For k = p, p-1, …..1:
     1. Consider all k models that contain all but one of the predictors in M_k for a total of k-1 predictors
     2. Choose the best among these models that yields the smallest RSS or largest R^2 and call it M_{k-1}
  3. Select the best model from $M_{0, \ldots\ldots\ldots} M_p$ using Cp, Akaike Information Criterion (AIC), BIC or adjusted R^2.

Backward Stepwise Selection

- Estimates 1 + p(p+1)/2 models which is a substantial improvement over best subset selection
- No guarantee that this approach yields the best model containing a subset of the p predictors
- Unlike forward selection, this method requires the number of observations is greater than the number of variables so that a full model can be fit.

Hybrid Approaches

- Hybrid approaches which are a combination of backward and forward approaches are also used to closely mimic the best subset selection while retaining the computational advantages of stepwise selection approaches

Choosing an Optimal Model

- R2 and RSS are related to training error. A model with all the variables will result in lower R2 and RSS. Our goal is to choose a model with low test error which needs to be estimated
- Two approaches to estimate test error
  - Indirectly estimate test error by making adjustment to training error to account for bias due to overfitting
  - Directly estimate test error using a validation set or cross validation

Choosing an Optimal Model

- MSE = RSS/n
- Training set MSE is an underestimate of test MSE
  - Regression coefficients are estimated to ensure small RSS on the training dataset
  - As more variables are included in the model, training error will decrease but test error may not
  - Hence training set R^2 and RSS cannot be used to compare models with different number of variables.
- A number of techniques that adjust for the model size are used to compare models with different numbers of variables. These techniques are discussed next

Cp

- The Cp estimate of the test MSE for a fitted least squares model with d predictors is calculated as

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

- $\hat{\sigma}^2$ is an estimate of the variance of the error ε associated with each response measurement
- Cp statistic adds a penalty of $2d\hat{\sigma}^2$ to the training RSS to adjust for the additional predictors and the tendency of training error to underestimate test error
- A model with low Cp is preferable since Cp takes on small values when test error is low.

Akaike Information Criterion (AIC)

- AIC is defined for large class of models fit by maximum likelihood. For a regression model with normally distributed errors, AIC is calculated as

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

- For least squares models, Cp and AIC are proportional and AIC offers no benefit

BIC

- Like AIC but derived from a Bayesian point of view. For a least squares model with d predictors, BIC is calculated as follows (excluding few irrelevant constants)

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

- A model with low BIC is preferable since it takes on small values when test error is low.

Adjusted R^2

- Another popular approach to compare models with different number of variables.

$$R^2 = 1 - \frac{RSS}{TSS}$$

- Adjusted R^2 for a model with d predictors is given by

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

- A large value of adjusted R^2 indicates a model with a small test error

Cross Validation

- Estimate the test error using cross validation for each model under consideration and choose the model for which the test error is minimum.
- Advantage over using AIC, BIC, Cp as it provides a direct estimate of the test error and makes fewer assumptions about the underlying model
- In the past, performing cross validation was computationally expensive which is no longer the case