7 Variable selection

A. Evaluating each potential subset of predictors

1. R^2-Adjusted

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

it is a measure of proportion of the variation in the observed response that is explained by the model. This measure will never decrease when adding extra predictors in the model. To penalize the complexity of the model, the usual practice is to choose the subset of the predictors that has the highest value adjusted $R^2 = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$. This is equivalent to choosing the subset of the predictors with the lowest value of $S^2 = RSS/(n-p-1)$ where p is the number of predictors in the subset. Choosing the subset of the predictors that has the highest value of adjusted R^2 tends towards over-fitting. For example, suppose that the maximum value is $R^2_{adj} = 0.692$ for a subset of $p = 10$ predictors, $R^2_{adj} = 0.691$ for a subset of $p = 9$ predictors . Even though R^2_{adj} increases when we go from 9 to 10 predictors there is very little improvement in fit and so the nine-predictor subset is generally preferred.

2.AIC and BIC

Suppose y_i,x_{i1},...,x_{ip} Maximizing likelihood function

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}RSS$$

is equivalent to minimizing

$$RSS = \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ...\hat{\beta}_p x_{ip}))^2$$

MLE of σ^2, $\sigma^2_{MLE} = \frac{RSS}{n}$, which is slightly different from our usual estimate $S^2 = \frac{RSS}{n-p-1}$. Estimated likelihood

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log(\frac{RSS}{n})$$

$$AIC = -2log(L(\hat{\beta}_0, ...\hat{\beta}_p, \hat{\sigma}^2)|Y)) + 2(p+2) = n\log(\frac{RSS}{n}) + 2p + \text{other terms}$$

Please note the other terms remain same for different subsets of predictors so R calculates

$$AIC = n\log(\frac{RSS}{n}) + 2p$$

to compare different subsets of predictors.

Bias Corrected version of AIC,

$$AIC_C = AIC + \frac{2(p+2)(p+3)}{n-p-1}$$

AIC_C should be used instead of AIC unless $\frac{n}{p+2} > 40$.Furthermore they recommend that AIC_C be used in practice since as n gets large AIC_C converges to AIC.

$$BIC = -2log(L(\hat{\beta}_0, ...\hat{\beta}_p, \hat{\sigma}^2|Y) + (p+2)\log(n)$$

When $n \geq 8$, $\log(n) \geq 2$ and the penalty term in BIC is greater than the penalty term in AIC.

A popular data analysis strategy which we shall adopt is to calculate R^2_{adj}, AIC, AIC_C and BIC and compare the models which minimize AIC, AIC_C and BIC with the model maximize R^2_{adj}.

B. Deciding on the collection of potential subsets of predictor variables

1. All possible subsets

Consider all 2^m possible regression equations and identify the subset that maximize a measure of fit or minimize an information criterion based on a monotone function of the residual sum of squares. Example bridge.txt. The best model has two or three predictors. Notice both of the predictors are judged statistically significant while just one predictor is judged statistically significant in the three-predictor model. Note p-values obtained after variable selection are much smaller than their true values. Two-predictor model is preferred.

Stepwise selection

Backward elimination, Forward selection bridge.txt

An important caution associated with variable selection

1. Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators that can not be ignored.

2. As a consequence, naive use of inference procedures that do not take into account the model selection step can be highly misleading.

Assessing the predictive ability of regression models

To assess the predictive ability of different regression models is to evaluate their performance on a new data set. In practice, this is often achieved by randomly splitting data into:

1. A training data set

2. A test data set.

The training data set is used to develop a number of regression models, while the test data set is used to evaluate the performance of these models.

Example: prostate cancer-prostateTraining.txt and prostateTest.txt.

Apart from a hint of decreasing error variance, the diagnostic plots further confirm that the model with 8 predictors without transformation is a valid model for the data. The dashed vertical line in the bottom right-hand plot is the usual cut-off for declaring a point of high leverage. Thus there are no bad leverage points. Notice that the overall F-test for model is highly statistically significant and four of the estimated regression coefficients are statistically significant (i.e. lcavol, lweight, lbph, svi). Finally we show the added variable plots associated with the model. Notice p=7 is judged as best when AIC is used, while AICC judges p=4 and BIC judges p=2. One could choose p=4 to avoid over-fitting. based on the output for models with 2, 4, 7 predictors, the model with 2-predictors seems to be preferred.

Model comparison using the test data set

Based on test data, none of these models is very convincing. The situation is due to

1. case 45 in the training set accounts for most of the statistical significance of the predictor lweight.

2. Splitting the data into a training set and a test set by randomly assigning cases does not always work well.

1. Remove case 45 from the training data. Note the dramatic effect on the optimal subsets of predictors. Thorough investigation is called for on the case 45.

2. Split the data into a training set and a test set such that the ”two sets cover approximately the same region and have the same statistical properties”.

Case 9 and 45 need further investigation.

Case 45 in the Training Set

We reconsider variable selection in this example by identifying the subset of the predictors of a given size that maximizes adjusted R-squared (i.e., minimizes RSS) for the training data set with and without case 45. Notice how the optimal two-, three- and five variable models change with the omission of just case 45. Thus, case 45 has a dramatic effect on variable selection. It goes without saying that case 45 in the training set should be thoroughly investigated. In summary, case 45 in the training data and case 9 in the test data need to be thoroughly investigated before any further statistical analyses are performed. This example once again illustrates the importance of carefully examining any regression fit in order to determine outliers and influential points. If cases 9 and 45 are found to be valid data points and not associated with special cases, then a possible way forward is to use variable selection techniques based on robust regression – see Maronna, Martin and Yohai (2006, Chapter 5) for further details. **LASSO**-least absolute shrinkage and selection operator, which we shall discover is a method that effectively performs variable selection and regression coefficient estimation simultaneously.

$$\min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{ip}))^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

Using a Lagrange multiplier argument, it can be shown that this is equivalent to

$$\min \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{ip}))^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

for some $\lambda > 0$. When $\lambda = 0$, it is equivalent to least square estimates. When λ is large (s is very small), some of the resulting coefficients are exactly zero, effectively omitting predictor variables from the fitted model. *BIC* is recommended to find the optimal LASSO model when sparsity model is of primary concern.