1

STAT2008/STAT6038

Model Selection

Evaluating Candidate Models

2

As the number of predictors increases, so do the number of candidate models to evaluate

If M is the number of predictors in a model, then 2^M is the number of candidate models

If $M = 10$, the number of candidate models is 1024

When $M = 20$, the number increases to 1048576.
That's a lot of models to evaluate!

What is a "good" model?

3

Usually we wish to identify a few "good" models.

How "good" a model is often depends on the purpose of the model.

For instance, we may want to learn about some aspect of the system from which the data are taken.

We may be interested in the sign of a particular coefficient.

We may be interested in which predictors are the "most important" when it comes to explaining the variation in the response

We may be interested in making forecasts for financial market variables (Here parsimony is important. We must take great care not to "overfit" our data and end up with a model that is describing a particular data set rather than the relationships between the predictors and the response)

Standard Model Comparison Criteria.

4

One way of comparing two different models is to compare the MSE from each of them and choose the model which has the smaller value.

Recall that if a model is "underspecified"; that is, we fit the model

$$Y = X_{(1)}\beta_{(1)} + \epsilon,$$

When the true model is:

$$Y = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \epsilon,$$

then the MSE of our regression will be an over-estimate of the true value of σ^2.

Adding predictors and MSE

5

Suppose we fit a "full" regression containing P parameters and a "reduced" regression containing $p < P$ parameters, and calculate the MSE's for these two regressions as s_P^2 and s_p^2, respectively.

If there were a large underspecification associated with the smaller model, then we would expect to see $s_P^2 < s_p^2$

We would tend to prefer the full model as more appropriate.

Note that as the number of parameters, P, in the full model nears the sample size, n, then s_P^2 will necessarily decrease down towards zero (regardless of the appropriateness of the model fit).

Parsimony

6

If the two estimates of σ^2 are nearly equal, or if $s_P^2 > s_p^2$, then we can conclude that the "reduced" model is the more appropriate of the two models.

Here we assume that we are comparing two *nested* models.

The same comparisons are possible between two non-nested models, however, the justification for choosing the model with the smaller mean square error is less clear

Coefficient of Determination

7

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

which measured the proportion of response variation "explained" by the regression, and could be used to compare nested and non-nested models

But this measure is not very useful for comparing across models, since the addition of any predictor to the model (whether or not it has any relationship to the response) will necessarily increase the value of R^2.

If we fit an n^{th} degree polynomial to a dataset of n pairs (x_i, Y_i), the R^2 value will be exactly 1.

Adjusted R squared

8

The *adjusted* R^2, or R^2_a, for a particular regression is defined by replacing the sums of squares in the R^2 definition with their corresponding mean squares, thus accounting for the number of parameters in the model:

$$R^2_a = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{s^2_e(n-1)}{SST}.$$

Clearly, however, there is a strong connection between the two measures which is demonstrated by the relationship:

$$R^2_a = \left(\frac{n-1}{n-p}\right)R^2 - \frac{p-1}{n-p}.$$

Cross-Validation

9

Initially, we might think that simply examining the ordinary residuals, e_i, will indicate how well our model will perform.

The problem with this approach is that the two components from which the residual is calculated, Y_i and $\hat{Y}_i$, are not "independent" of each other.

One general solution to this problem is to *split* our dataset, $S = M \cup V$ (with sample size n), into two groups, using the first (the *fitting sample* or *modelling set*, M, containing n_1 data points) and (the *validation sample or validation set*, V, containing the remaining $n_2 = n - n_1$ data points).

Cross Validation

10

We can evaluate the predictive performance by calculating a measure of discrepancy on the validation set, such as:

$$\sum_{i=n_1 \in V} (Y_i - \hat{Y}_i)^2 \quad \text{or} \quad \sum_{i=n_1 \in V} |Y_i - \hat{Y}_i|,$$

where the summation is over the indices for those data points in the validation set, V, and $\hat{Y}_i = x_i^T b_M$ where b_M is the least-squares estimator of the parameters of the model fit on the modelling set, M. We might then choose the model which has the smallest discrepancy measure as our preferred model.

How big should the data sets be?

11

There should be enough data points in the fitting sample to ensure adequate information for a regression analysis to make a "reliable" fit to the data.

The final regression estimates and predicted values should be based on the *entire* data set

That is, don't waste information!

If we have small n

12

In smaller datasets, data splitting may not be practical.

We can instead use $PRESS$ residuals,

$$e_{i,-i} = Y_i - \hat{Y}_{i,-i} = Y_i - x_i^T b_{-i} = \frac{e_i}{1 - h_{ii}}.$$

Note that the i^{th} $PRESS$ residual is the measure of discrepancy we would arrive at if we performed a data splitting cross-validation where the validation set consisted of solely the i^{th} data point. The $PRESS$ statistic, sometimes denoted as $PRESS_p$, is defined as:

$$PRESS_p = \sum_{i=1}^{n} e^2_{i,-i} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}}\right)^2.$$

PRESS

13

We may prefer the model or models with the smallest values of $PRESS_p$.

In addition, note that the $PRESS$ statistic is just a sum of squares for the $PRESS$ residuals.

So we might expect that its value would be close to the SSE if the candidate model under investigation was appropriate for the data.

A comparison between the $PRESS_p$ and SSE values can yield useful information regarding the appropriateness of a particular model.

Underfitting

14

If we fit an "underspecified" model to our data, we will incur a bias in our estimate of the error variance, σ^2.

If we fit the model $Y = X_1\beta_{(1)} + \epsilon$ when the true mean of the response variable is $E(Y) = X_1\beta_{(1)} + X_2\beta_{(2)}$, then it can be shown that the expected value of the MSE is:

$$E(s_p^2) = \sigma^2 + \frac{1}{n-p}\sum_{i=1}^{n}\{E(\hat{Y}_i) - E(Y_i)\}^2 = \sigma^2 + \frac{1}{n-p}\sum_{i=1}^{n}\{Bias(\hat{Y}_i)\}^2,$$

where p is the number of parameters in the underspecified model and $\hat{Y}_i$ is the fitted value for the i^{th} data point from the underspecified model.

The expectations are taken with respect to the true underlying model.

Bias

15

We will incur a bias in other estimates as well. For instance, the least-squares estimates from the underspecified regression have expectation:

$$\begin{aligned} E(b_{(1)}) &= E\{(X_1^TX_1)^{-1}X_1^TY\} \\ &= (X_1^TX_1)^{-1}X_1^TE(Y) \\ &= (X_1^TX_1)^{-1}X_1^T(X_1\beta_{(1)} + X_2\beta_{(2)}) \\ &= \beta_{(1)} + (X_1^TX_1)^{-1}X_1^TX_2\beta_{(2)} \\ &= \beta_{(1)} + A\beta_{(2)}, \end{aligned}$$

where $A = (X_1^TX_1)^{-1}X_1^TX_2$ is sometimes referred to as the *alias matrix*.

We can see that the least-squares estimates from an underspecified model will be biased, which implies that any predictions will also be biased.

Overfitting

16

If we "overspecify" our model, then the variances of the least-squares estimates are inflated.

In other words, suppose that the true model for our response is $Y = X\beta + \epsilon$ and that b represents the least-squares estimates from a regression of Y using the design matrix X.

If we were to fit the model $Y = X\beta + \beta_{k+1}x_{k+1} + \epsilon$, then the new least-squares estimator, b^*, would have the property:

$$Var(b_j^*) \geq Var(b_j) \qquad \text{for } (j = 1, \ldots, k).$$

Overfitting

17

The variance of predicted values will also be inflated.

In other words, at a particular set of predictor values $x_0 = (x_{1,0}, \ldots, x_{k,0}, x_{k+1,0})$, we would define the predicted values from the two regressions noted above by

$$\hat{Y}(x_0) = \sum_{i=1}^{k} b_i x_{i,0} \qquad \text{and} \qquad \hat{Y}^*(x_0) \sum_{i=1}^{k+1} b_i^* x_{i,0},$$

respectively, and the variances of these predictions would satisfy:

$$Var\{\hat{Y}(x_0)\} \leq Var\{\hat{Y}^*(x_0)\}.$$

Overfitting

18

The MSE from the overspecified model will still be an unbiased estimate of σ^2, so that

$$E(s_{k+1}^2) = \sigma^2.$$

However, it will be a less precise estimate than s_p^2, the MSE from the correct model. This is because it is now an estimator based on one fewer degree of freedom.

Bias/Variance Trade off

19

Model selection amounts to finding an appropriate compromise between the bias of an underspecified (or *underfit*) and the inflated variances of an overspecified (or *overfit*) model.

To help decide on a reasonable compromise between these two poles, we need a criterion which is sensitive to the discrepancies inherent in both under- and overfitting.

Mean Squared Error of Prediction

20

Consider a measure based on how well the particular model under investigation predicts, such as the *mean squared error of prediction* at a point x_0:

$$MSE\{\hat{Y}(x_0)\} = Var\{\hat{Y}(x_0)\} + \left[E\{\hat{Y}(x_0)\} - E\{Y(x_0)\}\right]^2$$

$$= Var\{\hat{Y}(x_0)\} + \left[Bias\{\hat{Y}(x_0)\}\right]^2$$

which clearly incorporates both of the issues at hand.

Scaled Mean Squared Errors

21

Unfortunately, this measure depends upon the particular values, x_0, we choose for the predictor variables.

To overcome this problem, we could look at the scaled mean squared errors at each of the fitted values:

$$\sum_{i=1}^{n} \frac{MSE\{\hat{Y}_i\}}{\sigma^2} = \sum_{i=1}^{n} \frac{Var\{\hat{Y}_i\} + \left[Bias\{\hat{Y}_i\}\right]^2}{\sigma^2}.$$

The above quantity will not reflect the interpolation or extrapolation capabilities of the candidate model.

It will however give us a quantity that can be used to obtain a workable balance between bias and inflated variances.

22

A little algebra will show that if the candidate model has p parameters, then

$$\sum_{i=1}^{n} \frac{Var\{\hat{Y}_i\}}{\sigma^2} = p,$$

which follows using a nearly identical argument to that used to demonstrate that the sum of the leverages, $\sum_{i=1}^{n} h_{ii}$, is equal to the number of parameters in the model (which is also the number of columns in the design matrix $X_{(1)}$).

Mallows' C_p statistic

23

Similarly, using the previous result regarding the expectation of the MSE of an underspecified model, we can see that:

$$\sum_{i=1}^{n} \frac{\left[Bias\{\hat{Y}(x_0)\}\right]^2}{\sigma^2} = \frac{(n-p)(E(s_p^2) - \sigma^2)}{\sigma^2}$$

$$\approx \frac{(n-p)(s_p^2 - \sigma^2)}{\sigma^2},$$

where s_p^2 is the MSE from the candidate model under investigation. So, if we had an "independent" estimate of σ^2, say $\hat{\sigma}^2$, then we could estimate the sum of the scaled mean squared errors of the fitted values using *Mallows' C_p* statistic:

$$C_p = p + \frac{(n-p)(s^2 - \hat{\sigma}^2)}{\hat{\sigma}^2}.$$

24

We might then favour the model or models with the smallest values of C_p. Also, note that a potentially reasonable norm by which to judge the size of the C_p value can be based on the fact that a model with no bias would have $C_p = p$.

We don't generally have an "independent" estimate of the true error variance.

It is often suggested that a reasonable choice for $\hat{\sigma}^2$ is the MSE from the "full" model, s_F^2, derived using all the predictors under consideration (recall that the MSE from an overspecified model is still an unbiased estimate of σ^2, although it is somewhat less precise than the MSE calculated from the true model).

25

This choice for $\hat{\sigma}^2$ means that the C_p value for the "full" model will necessarily be exactly equal to its total number of parameters.

Frequently, a plot of the C_p values versus the associated number of parameters, p, for each of the candidate models is used to visually display the information contained in this selection criterion. For reference, the line $C_p = p$ is also generally superimposed on the plot.