Model Selection Criteria and Automated Search Procedures

1 Introduction

In building a regression model, we are faced with two conflicting objectives:

1. to include more predictors into the model so as to improve the predictive ability of the model and
2. to not include predictors that are unnecessary, which will lead to more uncertainty in our predictions and make our model needlessly complicated, making predictions and interpretations more challenging.

For example, recall the NFL dataset that you have seen and worked on in the previous modules. In the data set, you are presented with 9 potential predictors to predict the number of wins for a team. You are faced with the question of which and how many of the predictors you will need to include in the model. Do we start by using just offensive statistics, or should we start by also including the strength of schedule? Or why not just use all of the predictors right away?

In this module, you will learn the various ways to assess different competing models.

2 Model Assessment Concepts

Previously, we looked at **model diagnostics**, which are tools to assess if the assumptions for a regression model are met. We are assessing if the assumption that the error terms in a regression model and independent and identically distributed as a normal distribution with mean 0 and constant variance denoted by σ^2.

Meeting the assumptions do not guarantee that the model fits the data well, as the variance of the error terms could be large. We now shift our attention to **model assessment**, which measures how well the model fits the data and how well the model predicts future observations.

Recall there are two main uses of regression models:

1. **Prediction**: Predict a future value of a response variable, using information from predictor variables.
2. **Association**: Quantify the relationship between variables. How does a change in a predictor variable change the value of the response variable, on average?

With multiple predictors to choose from, we need a way to chose between all the possible models that we can fit. For example, if we have k predictors, then there are 2^k possible models with just additive effects to consider. What predictors should we include?

- Generally speaking, we can improve the model fit (in terms of improving R^2 or reducing SS_{res}) by adding more predictors (even if added needlessly).
- However, if we needlessly add in predictors, the predictive ability of a model can suffer, and we make the model more difficult to interpret.

So we can see that needlessly adding predictors negatively affects both uses of regression models. Generally, we apply the principle of "Ockham's razor": Given two theories that describe a phenomenon equally well, we should prefer the theory that is simpler. In selecting regression models, we should choose the one with fewer parameters when there are multiple models that give nearly the same fit to the data. We always should be asking: does the improvement in fit for models with more parameters justify the extra complexity?

Model assessment can be viewed as a trade-off between model complexity and model fit, so the model can be interpreted and predicts well on future data.

2.1 Training and test data

By now, you should realize that a model that fits the data well does not necessarily predict well on future observations. We introduce the definitions of training and test data:

- **Training data**: are the observations used to build the regression model.
- **Test data**: are the observations that were not used to build the regression model and are solely used to assess how well the model predicts future observations.

Model fit typically measures how well the training data fits the model (e.g. R^2, SS_{res}). Model assessment measures how well the model does in predicting the test data.

2.2 Overfitting

A model that fits the training data well is not guaranteed to predict test data. In fact, a model that performs a lot worse on test data than on training data is an indication that the model is needlessly complicated. Such a model is **overfitted**.

It may seem counter intuitive that a model that has more parameters could perform worse on test data. The reason is the following: we have mentioned that models take the form $y = f(\boldsymbol{x}) + \epsilon$, where the function f denotes the true relationship between the response and predictor variables, and ϵ denotes the random error. A model that is overfitted fails to account for the errors properly by incorporating the errors into the estimation of f, so the estimated f is not a good approximation for the true relationship f. So we end up using this poor approximation for f in predicting test data.

Model selection criteria are used to select between several models by assessing the models in terms of model fit and model complexity. These criteria prevent overfitting from happening.

3 Model Selection Criteria

A variety of model selection criteria have been proposed to assess regression models. These criteria typically measure the model fit and have a penalty for each additional parameter. The penalty penalizes the model when it gets too complicated.

The coefficient of determination, R^2, is typically not used as a model selection criteria, since it always increases when we add parameters to the model, and so will favor more models that add parameters. R^2 does not have a penalty for each additional parameter. R^2 can be used when comparing models with the same number of parameters.

We will next look at various model selection criteria that have a penalty for each additional parameter.

3.1 Adjusted R-squared: $R^2_{Adj,p}$

The adjusted R-squared, denoted by $R^2_{Adj,p}$, is

$$R^2_{Adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p) \tag{1}$$

- $R^2_{Adj,p}$ is a measure of fit with penalty for additional parameters.
- By including one additional parameter, R^2_p will increase and the divisor $n - p$ decreases.
- Choosing the model with the largest $R^2_{Adj,p}$ is equivalent to selecting the one with the smallest $MS_{res}(p)$.
- Larger value is better.

3.2 Mallow's C_p

Mallow's C_p measures the variance and bias associated with predicting test data. In the context of MLR, it is found using

$$C_p = \frac{SS_{res}(p)}{MS_{res}(P)} - (n - 2p). \tag{2}$$

- C_p is a measure of fit with penalty for additional parameters.
- By including one additional parameter, $2p$ increases while SS_{res} decreases.
- $MS_{res}(P)$ denotes the MS_{res} of the model with all P parameters.
- Some authors suggest choosing the simplest model with C_p closest to p. The argument is that if the model is unbiased, $C_p = p$.
- However, unbiased models are not guaranteed to perform best on test data. So, we should select the model with the smallest C_p.

3.3 AIC_p and BIC_p

A couple of related measures are also used, the Akaike information criterion, AIC_p, and the Bayesian information criterion, BIC_p. In MLR, they are

$$\text{AIC}_p = n \log \frac{SS_{res}(p)}{n} + 2p \tag{3}$$

and

$$\text{BIC}_p = n \log \frac{SS_{res}(p)}{n} + p \log n. \tag{4}$$

- Both are measures of fit with penalty for additional parameters.
- By including one additional parameter, $2p$ and $p \log n$ increase while SS_{res} decreases.
- Models with low AIC_p, BIC_p are desired.
- If $\log n > 2$ (i.e. $n \geq 8$), then BIC_p increases with p more quickly than AIC_p. Thus, BIC_p favors smaller models compared to AIC.

3.4 PRESS statistic

The PRESS statistic measures the difference in predicted response when an observation is excluded in estimating the model. It is defined as

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i(i)}\right)^2 \tag{5}$$

where $\hat{y}_{i(i)}$ is the predicted value for the ith observation when the regression is fitted without the ith observation.

- Small PRESS statistic is desired.
- Models with small PRESS statistics have small prediction errors on test data.
- Of the criteria listed, the PRESS statistic (5) is motivated by measuring the prediction error in test data, whereas the other criteria (1), (2), (3), (4) are motivated by balancing model fit and model complexity, and as a consequence will perform well in predicting test data.

A related measure to the PRESS statistic (5) is $R^2_{prediction}$,

$$R^2_{prediction} = 1 - \frac{PRESS}{SS_T}, \tag{6}$$

which can be interpreted as the **proportion of the variability in the response of new observations that can be explained by our model**.

- On average, $R^2_{prediction}$ is less than R^2_p.

- An $R^2_{prediction}$ that is a lot smaller than R^2_p indicates overfitting.

3.5 Summary and final comments

We have introduced a number of criteria. While these criteria attempt to measure model fit and complexity, they are not mathematically equivalent. None of these should be regarded as definitive in itself, but taken together they provide a range of possible options which, combined with some judgment about the data themselves, can usually be used to decide upon a model which is reasonable from all criteria. A couple of other points to note:

- We still need to check the model for regression assumptions.
- These criteria can only be used to compare models with the same response variable, as they are affected by the unit associated with the response variable.

4 Automated Search Procedures

With multiple predictors to choose from, we need a way to choose between all the possible models that we can fit. For example, if we have k quantitative predictors, then there are 2^k possible models with just additive effects to consider. If k is large, then there are many models to compare. Can we automate the process of model selection to make things more computationally efficient?

4.1 Forward selection, backward elimination

A couple of methods to automate this process:

1. **Forward Selection**:
   - We begin with a model with no predictors, and add in predictor variables one at a time in some optimal way, until a desirable stopping point is reached.
2. **Backward Elimination**:
   - We begin with a model with all potential predictors, and then remove the "weak" predictor variables one at a time in some optimal way, until a desirable stopping point is reached.

A critical concept is that each step is conditional on the previous step. For instance, in forward selection we are adding a variable to those already selected.

The criterion to add (or eliminate) a predictor in each step for forward selection (or backward elimination) might be based on p-value, SS_{res}, AIC_p, etc.

Suppose we use AIC_p as the criterion. We know that smaller values of AIC_p are desired.

- In a forward step, given a model of size p in the previous step, compare all the candidate models of size $p+1$ by adding one of the remaining predictor variables. Among these size-$(p+1)$ models, select the one that results in the smallest AIC_{p+1}, which is also smaller than the AIC_p of the model in the previous step.
- In a backward step, given a model of size p in the previous step, compare all the candidate models of size $p-1$ by eliminating one of the existing predictors. Among these size-$(p-1)$ models, remove the one that results in the smallest AIC_{p-1} value, that is also smaller than the AIC_p of the model in the previous step.
- The algorithm stops when the AIC no longer decreases.

Let us consider this toy example. Suppose there are four potential predictors, x_1, x_2, x_3, x_4. The table below gives the AIC of all possible models. Assume the AIC of the intercept-only model is 25.

Model	AIC	Model	AIC	Model	AIC	Model	AIC
x_1	20	x_1, x_2	8	x_1, x_2, x_3	5	x_1, x_2, x_3, x_4	10
x_2	17	x_1, x_3	10	x_1, x_2, x_4	7		
x_3	15	x_1, x_4	17	x_1, x_3, x_4	8		
x_4	19	x_2, x_3	11	x_2, x_3, x_4	6		
		x_2, x_4	12				
		x_3, x_4	9				

Suppose we start from the intercept-only model. What model gets selected by forward election?

- In step 1, we will add x_3 to the intercept-only model. This is the model with just 1 predictor that has the smallest AIC that also decreases the AIC from the previous step.
- In step 2, we consider adding one of the three remaining predictors, x_1, x_2, x_4 to the model with x_3 already in. x_3 will not be removed. x_4 will be added because it results in the smallest AIC that also decreases the AIC from the previous step.
- In step 3, we consider adding one of the two remaining predictors, x_1, x_2 to the model with x_3, x_4 already in. x_2 will be added because it results in the smallest AIC that also decreases the AIC from the previous step.
- Algorithm stops here, as adding an additional predictor at this step will not decrease the AIC. So the model selected has x_2, x_3, x_4.

4.1.1 Practice question

What model gets selected by backward elimination, if we start from the model with all 4 predictors?

View the associated video for a review of this practice question.

4.1.2 Stepwise regression

A combination procedure called **stepwise regression** can also be used. It is a combination procedure because at each step, the algorithm considers adding any of the remaining predictors or removing any of the current predictors.

4.2 Final comments

We still need to check the model for regression assumptions. Also, the model chosen by these procedures are not guaranteed to be the same. The starting point, and criteria used, can also affect the model selected. Typically, automated search procedures consider models with just additive effects.

Use these procedures as a starting point in model building, and not as a definitive answer to choose your final model.