Section 11: Model Selection & Comparison

TF: Reagan Rose (rrose@g.harvard.edu)

December 2, 2015

Model Selection

When we conduct simple linear regression and perform a t-test to determine whether $\hat{\beta}_1$ is statistically significant, what we are really doing is asking a question of **model selection**. What we are asking is the following:

Does the model with a linear relationshop between X and Y explain the data significantly better than a model with just an intercept?

In essence, the problem of **model selection** in regression consists of the following:

- Which predictor variables should be included in the model?
- What power of each predictor variable must be included?
- Are interaction terms between any variables necessary?

In theory, if we have a dataset with n observations and p predictors, we would want to use all p predictors to fit a model to the data, since chances are that each of the predictors will contribute at least a small bit to explaining the variability of the outcome variable Y (though remember, with the methods we have, we cannot run regression when $p > n$). However, this can lead to a serious problem called **overfitting**. In general, with model selection problems, we are faced with two competing forces:

- **In-sample fit:** We want as many predictors as possible, as this yields better and better fit for the observed data
- **Out-of-sample fit:** Having arbitrarily many predictors leads to overfitting, and therefore reduced performance on new data, which is ultimately the purpose of the model

To compromise, we seek **model parsimony**: that is, to retain as much explanatory power as possible (measured in terms of R^2, reduced SSE, etc.) while keeping the parameters to the minimum number.

There are two main methods to (manually or automatically) conduct model selection (besides using adjusted R^2 or information criteria):

- **Forward selection:** Start with an empty or basic model. Add main effects terms based on domain knowledge and assumed importance, keeping only those that have significant p-values. Add interaction terms for main effects terms added to the model, again with adequately low p-values.
- **Backward elimination:** Start with every main effect and interaction term. Eliminate any interaction terms with p-value $> \alpha$, and re-run the regression. Continue whittling down terms; whenever a main effect term is eliminated, any interaction terms with the main effects term must also be eliminated.

In any model selection procedure, there are two points to keep in mind:

- Ultimately, model selection should not take precedence over domain knowledge. How you start tje model selection process and deciding which terms to keep is an art, with a fair share of analyzing p-values tempered with your intuition for how important and interpretable a term is.
- When comparing R^2 values, note that adding terms will **always** increase R^2 (refer to Section 10 notes); thus, one must only be concerned with *how much of an increase* in R^2 is provided by adding another term.

Information Criteria

Akaike's Information Criterion (AIC)

The basic idea behind AIC is that if we knew the model and distribution of Y (i.e. with given predictors), then we could compare two models M_0 and M_1 by comparing how much information is *lost* by using M_0 versus using M_1. Unfortunately, we don't know the true model/distribution of Y; that's what we're trying to approximate with model selection! However, one can still compare two models by this method. We define

$$AIC = nlog(SSE/n) + 2p$$

where l is the log-likelihood of the Normal distribution on Y.

Note that AIC penalizes the number of parameters as a constant term; that is, for a given number of parameters, a model's penalty does not scale with the number of observations.

Bayesian Information Criterion (BIC)

An alternative information criterion is the BIC, defined by

$$BIC = nlog(SSE/n) + plog(n).$$

The crucial difference between BIC and AIC is that BIC penalizes the number of parameters more heavily, but scales with the number of observations (this is the idea of *letting the data speak for itself*).

Cross Validation

A method mostly developed by and adopted from machine learning is a method of approximating the "sweet spot" of optimality. The main problem of overfitting arises because we are considering only the observed data, and therefore have no way of measuring how our model would do on unobserved data. Thus, one trick is to consider only a *part* of our data as observed, isolating another part as "unobserved" in order to test how well the model fits. Cross validation runs as follows:

1. Partition the data into a **training** and **test** set.
2. **Fit the model** using the training set.
3. Conduct **model selection** using the test set.
4. Repeat using different partitions, and either select the best or average over results.

Multicollinearity

This isn't related *exactly* the the topic of model comparison, but it is important so I want to briefly touch on it. Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated. When it exists, it can cause major problems in an analysis and thereby severely liit the research conclusions we can draw. When multicollinearity exists. we face the risk of experiencing the following problems:

- the estimated regression coefficient of one variable depends on which other predictors are included in the model
- the precision of the regression coefficients decrease as more predictors are added to the model
- the marginal contribution of any one predictor variable in reducing RSS depends on which other variables are already in the model
- hypothesis tests for $\beta_k = 0$ may be unreliable, depending on which predictors are in the model

Multicollinearity is often the result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which data are collected. In general, we cannot fix multicollinearity (methods exist, but are beyond the scope of this course), but in order to understand the *scope and limitations* of our analyses, we must check for it. If you find that you have strong multicollinearity in your data, you should **be wary about any conclusions you make** using your regression analysis.

Exercises

Question 1. Model Comparison & Hypothesis Tests
To understand the different methods we can use when doing model comparison, we're going back to the "InsuranceClaims.csv" dataset we've used before.

(a) Write down the regression equation for predicting claim amount using a full interaction model with claim type and coverage as predictors. Calculate the slope of coverage for each of the claim type groups. Which group has the smallest slope?

(b) Test the hypothesis that the slope for fire claims is different from the slope for theft claims using the "gamma" test for linear combinations of coefficients. Confirm your results in R.

(c) Now use the extra sum of squares test to determine whether the slopes for all different claim types are significantly different from each other.

(d) Using the same method, determine whether the interaction terms are significantly different from 0. That is, is the model including interaction terms better than the model without interaction terms?

(e) Calculate the AIC of 1) the first model, 2) the model without interaction terms, 3) the model with only coverage as a predictor, and 4) the model with no predictors. Which is best in terms of AIC?

Question 2. Model Selection & Cross Validation
We've just done an initial model comparison using AIC. Now let's try some more compact methods of doing model comparison.

(a) Using the same data and models as above, run a forward, backward, and stepwise regression procedure to select which of the 4 models is best.

(b) Now use a five-fold cross-validation approach to compare the four models. Which is best?

Question 3. Multicollinearity
Come up with a "toy example" to prove to yourself the problem of multicollinearity. To do this, generate three datasets: 1) where all variables are perfectly uncorrelated, 2) where all variables are only weakly correlated, and 3) where variables are perfectly correlated. In each case, run regressions with different predictors included in the model and see how the the regressions for dataset (3) behave erratically.