Stat 849: Subset selection and obtaining a sequence of models for model selection

There are, in general, two main reasons why we may not be "happy" with least squares estimators (LSEs) in a regression model. The first reason is the *prediction accuracy*, which reflects the fact that LSEs may be unbiased but have high variance. Prediction accuracy can sometimes be improved by shrinking (we will talk about this later in the class) or setting some coefficients to zero. With this, some bias is sacrificed to reduce the variance of the predicted values with the goal of improving overall accuracy. The second reason for considering alternatives to LSEs is the issue of interpretability. If we have thousands of variables (predictors), we might usually be interested in identifying a subset of these that show the strongest effects. (Reference: Hastie, Tibshirani, Friedman 2002).

So far, we have seen various criteria to choose among different models (C_p, AIC, BIC, adjusted-R^2). We have yet to talk about obtaining candidate models when we have large number of predictor.

Subset selection refers to only keeping a subset of the original variables, and eliminating the rest from the model. Least squares is used for estimation in the obtained smaller models.

Suppose that we have p predictors. There are alternative ways of obtaining sequence of models of various sizes using these predictors. We list these below:

1. **Best subset regression.** This is feasible to employ if one has a small subset of predictors. The general principle includes finding for each $k \in \{1, \cdots, p\}$ the subset of size k that gives smallest residual sum of squares.

   Question: Given the model size, e.g. $k = 3$, how do you decide which model of size k is the best? Note that if there are p predictors, we have p choose k models of size k.

   If we have a candidate model for each model size (note that each candidate model implies a candidate estimator $\mu_{n,k}$ for $E[Y]$), then the next step is to choose among these candidate estimators using the various approaches we talked about.

   Drawbacks: Can you think of any?

2. **Forward stepwise selection (Forward addition).** This methods starts with the intercept and sequentially adds into the model the predictor that improves the fit most. First note that this will give us a nested sequence of models with sizes from 1 to p.

   Suppose our current model has k predictors ($k+1$ parameters including the intercept term) and the corresponding parameters are represented by the vector $\hat{\beta}_k$. Say we add in a predictor resulting in estimates $\tilde{\beta}$. The improvement in the fit is often based on the F-statistics:

$$F = \frac{RSS(\hat{\beta}_k) - RSS(\tilde{\beta})}{RSS(\tilde{\beta})/(n-k-2)} \sim F_{1,n-k-2}.$$

   A commonly employed strategy adds in sequentially the predictor producing the largest value of F, stopping when no predictor produces an F-ratio greater than the 90th or 95th percentile of the corresponding distribution $F_{1,n-k-2}$.

   Drawbacks: Can you think of any?

3. **Backward stepwise selection (Backward deletion, backward elimination).** This procedure starts with the full model, and sequentially deletes predictors. Similar to forward selection, it typically uses an F-ratio to choose the predictor to delete. In this case, we drop the predictor producing the smallest value of F at each stage, stopping when each predictor in the model produces a value of F greater than the 90th or 95th percentile when dropped.

   Question: Can we use this method regardless of the dimensions of n and p?

   Drawbacks: Can you think of any?

4. **Stepwise regression (Forward selection - Backward elimination).** This is a combination of forward stepwise selection and backward stepwise elimination procedures, and in particular, consists of forward selection (FS) followed by backward deletion (BD) at each step. There are two model parameters F_{IN} and F_{OUT}.

   Start with a model consisting of the intercept term alone, then perform a FS step adding a single variable if the corresponding F value is greater than F_{IN}. Then, perform a BD step removing a variable if the corresponding F value is less than F_{OUT}. These steps are iterated until no further variables are introduced at the FS step. Provided that $F_{OUT} < F_{IN}$ algorithm must eventually terminate (see page 418 in Seber and Lee).

   Drawbacks: Can you think of any?

Question. How about including interactions?

Useful `R` functions: `leaps(), mle.stepwise(), step()`. See `R` handout for an example.

These procedures provide means for obtaining a set of candidate models (also estimators). Once we have those models we would like to be able to choose the model that provides the best *prediction accuracy* [Recall that the selection criteria we talked about do not utilize future observations]. Say as our estimate of $E[Y \mid X]$ we choose $\mu_{n,k}(X) = X_k\beta_{k,n}$, where $\beta_{k,n}$ is the LSE in a k dimensional model and X_k is the corresponding design matrix in this model. Then, we want to know what the *expected loss (expected squared error loss)* would be for future observations. In other words,

$$E_{P_0}[(Y - X\beta_{k,n})^2],$$

which represents the risk of the estimator $\beta_{k,n}$ is the quantity of interest. Here, the expectation is w.r.t. unknown true data generating distribution P_0. We need to find a way of estimating this quantity! Also note that the risk is a random quantity because it depends on the random sample (also called the *learning sample*) through $\beta_{k,n}$. We will use *cross-validation* to estimate this risk and then choose the estimator with minimum cross-validated risk.