Model Selection

There are two model selection procedures we will consider:

- all-possible regressions procedures identify "good" subsets of the pool of potential independent variables among all possible subsets of the variables, where "good" may be defined with respect to several criteria
- automatic search procedures (such as forward stepwise regression) that search for the "best" subset of independent variables without comparing all possible regressions

1. All-Possible-Regressions Procedures

The all-possible-regressions procedure examines all the 2^{P-1} possible subsets (of 1, 2, ..., P – 1 variables) of the pool of P – 1 potential X variables and identifies a few "good" subsets according to one of the criteria below. These criteria can also be used outside the all-possible-regressions context, to compare two or more regression models for the same dependent variable. (There are 2^{P-1} subsets because each of the P – 1 variables can be either included or excluded from a model.)

1. R_p^2 (or SSE_p) Criterion

With the R_p^2 criterion (where the p subscript refers to the number of variables in the model) subsets of the potential X variables for which the ordinary R-square is large are considered "good". Choosing the model with largest R^2 is equivalent to choosing the model with smallest SSE (since $R^2 = 1 - SSE/SSTO$ and SSTO is constant across all models). The R_p^2 criterion is used to judge when to stop adding more variables rather than finding the "best" model, since R_p^2 can never decrease when p increases.

2. R_a^2 (or MSE_p) Criterion

The R_a^2 criterion compares models on the basis of the adjusted R-square, which adjusts for the number of independent variables included. It can be shown that $R_a^2 = 1 - MSE/(SSTO/(n-1))$, so that maximizing R_a^2 is equivalent to minimizing MSE.

3. AIC and SBC Criteria

AIC (Akaike Information Criterion) is defined as
$AIC_p = n \ln(SSE_p) - n \ln(n) + 2p$

SBC (Schwatz's Bayesian Criterion) is defined as
$SBC_p = n \ln(SSE_p) - n \ln(n) + [\ln(n)]p$

For both criteria *smaller values* are better. Note that both criteria *increase* with SSE (poor model fit) and with *p* (number of independent variables). Thus both criteria penalize models with many independent variables.

2. Automatic search procedures

There are three common types of automatic search procedures:

1. stepwise regression
2. forward selection
3. backward elimination

See the SAS code for illustration of how these procedures work.

The major weakness of these techniques, compared to the all-possible-regressions methods, is that the end result is a single "best" model. This model may not be as desirable as other models missed by the procedure.

Collinearity

Sometimes "multicollinearity"

1. Symptoms of collinearity

The following symptoms may indicate a collinearity problem

- large changes in b_k when adding or deleting variables(s) or observation(s)
- b_k non significant for a theoretically important X_k
- b_k with sign opposite of expected (from theory or previous results)
- large correlations(s) in r_{XX}
- wide $s\{b_k\}$ for important X_k(s)
- b_ks are non significant even though F for whole regression is significant

2. Tolerance (TOL) & Variance Inflation Factor (VIF)

The tolerance for variable X_k is

$(TOL)_k = 1 - R_k^2 \quad k = 1, 2, ..., p-1$

where R_k^2 is the R square when X_k is regressed on the other independent variables in the model including a constant.

The variance inflation factor for variable X_k is the inverse of the tolerance
$(VIF)_k = 1/(TOL)_k$

Using TOL or VIF for Diagnosis

As discussed earlier, a common rule of thumb is to take

- TOL < .1 or equivalently
- VIF > 10

as an indication that collinearity may be a problem.

3. Remedies for Collinearity

The fundamental problem with collinearity is that the pattern of intercorrelations among variables makes the **X'X** matrix nearly singular, which makes estimation of the regression coefficients imprecise/unstable. Once collinearity has been diagnosed, a number of strategies may be considered, starting with the most obvious ones:

- collinearity does not affect the precision of the predicitons $\hat{Y}_h$ or $\hat{Y}_{h(new)}$ as long as X_h follows the same pattern of collinearity as the bulk of the data so if the main purpose of the analysis is prediction, collinearity is not an issue
- if collinearity is in a polynomial regression, center X by transforming X_i into $(X_i - X_.)$
- often collinearity arises when the Xs are conceptually related, alternative measures of the same theoretical concept; if so either
    - drop one or more of the collinear variables (keeping in mind the danger of specification bias)
    - incorporate collinear variables into a single index (by summing or averaging)
    - calculate one or several principal components on the subset of collinear variables and use the component(s) in the regression instead of the original variables
    - if estimating the separate effect of collinear variables that are conceptually related (i.e., alternative measures of the same concept) is not essential, one may be content to simply test their joint effect on the dependent variable
- in some cases a pattern of collinearity can be broken by collecting additional data; this is more often feasible in experimental than in observational studies
- if all else fails (i.e., you can't let go of these extra Xs) use *ridge regression*
    - Ridge regression introduces a small bias in order to reduce $s\{b_k\}$. The goal is an estimator that has a higher probability of being close to the true value of the coefficient.