SELECTING TERMS (Supplement to Section 11.5)

Transforming toward multivariate normality helped deal with the problem that deleting terms from the full model might result in a non-linear mean term or non-constant variance.

Another possible problem: Dropping terms might introduce bias.

First observe: When we drop terms and refit using least squares, the coefficient estimates may change. Example: The highway data.

Explanatory Example: Suppose the correct model has mean function \( E(Y | x) = \eta_0 + \eta_1 u_1 + \eta_2 u_2 + \epsilon \). (So \( \epsilon \) is a random variable with \( E(\epsilon) = 0 \).)

Suppose further that \( u_2 = 2u_1 + \delta \), where \( \delta \) is a random variable with \( E(\delta) = 0 \).

Then \( Y = \eta_0' + \eta_1' u_1 + \eta_2' (2u_1 + \delta) + \epsilon' \) where \( \eta_0' = \eta_0, \eta_1' = \eta_1 + 2\eta_2, \) and \( \epsilon' = \eta_2 \delta + \epsilon \). Since \( E(\epsilon') = E(\eta_2 \delta + \epsilon) = \eta_2 E(\delta) + E(\epsilon) = 0 \),

the mean function for the submodel is \( E(Y | x) = \eta_0' + \eta_1' u_1 \).

Now suppose we fit both models by least squares, giving fits \( \hat{y}_i \) for the full model and \( \hat{y}_{i_{sub}} \) for the submodel. Recalling that 1) the least squares estimates are unbiased for the model used, 2) \( u_{i1} \) denotes the value of term \( u_1 \) at observation \( i \), etc., and 3) we are fixing the \( x \)-values, and hence the \( u \)-values, of the observations, we have that the expected values of the sampling distributions of \( \hat{y}_i \) and \( \hat{y}_{i_{sub}} \) are:

\[
E(\hat{y}_i) = \eta_0 + \eta_1 u_{i1} + \eta_2 (2u_{i1} + \delta_i) \quad \text{where} \quad \delta_i \text{ is the value of } \delta \text{ for observation } i, \text{ and} \\
E(\hat{y}_{i_{sub}}) = \eta_0' + \eta_1' u_{i1} = \eta_0 + (\eta_1 + 2\eta_2) u_{i1}.
\]

Note that \( E(\hat{y}_i) \) has the additional term \( \eta_2 \delta_i \) that \( E(\hat{y}_{i_{sub}}) \) doesn’t have. Thus, if the full model is the true model, then \( \hat{y}_{i_{sub}} \) is a biased estimator of \( E(Y | x) \).

Definition: The bias of an estimator is the difference between the expected value of the estimator and the parameter being estimated. So for parameter \( E(Y | x) \) and estimator \( \hat{y}_{i_{sub}} \),

\[
bias(\hat{y}_{i_{sub}}) = E(\hat{y}_{i_{sub}}) - E(Y | x)
\]

A counterbalancing consideration: Dropping terms might also reduce the variance of the coefficient estimators -- which is desirable! To see this, we use a formula (see Section 10.1.5) for the sampling variance of the coefficient estimators: The variance of the coefficient estimator \( \hat{\eta}_j \) in a model is
\[ \text{Var}(\hat{\eta}_j) = \frac{\sigma^2}{SU_j U_j \left(1 - R_j^2\right)}, \]

where \( SU_j U_j \) is defined like \( SXX \), and \( R_j^2 \) is the coefficient of multiple determination for the regression of \( u_j \) on the other terms in the model. Notice that the first factor is independent of the other terms. Adding a term usually increases \( R_j^2 \); deleting one usually decreases \( R_j^2 \). Thus adding a term usually increases \( \text{Var}(\hat{\eta}_j) \); deleting a term usually decreases \( \text{Var}(\hat{\eta}_j) \) (i.e., gives a more precise estimate of \( \eta_j \)). Since \( \hat{y}_i \) is a linear combination of the \( \hat{\eta}_j \)'s, the effect will be the same for \( \text{Var}(\hat{y}_i) \).

**Summarizing:** Dropping terms might introduce bias (bad) but might reduce variance (good). Sometimes, having biased estimates is the lesser of two evils. The following picture illustrates this: One estimator has distribution \( N(0, 1) \) and is unbiased; the other has distribution \( N(0.5, 0.1) \) and is hence biased but has smaller variance:

One way to address this problem is to evaluate the model by a measure that includes both bias and variance. This is the **mean squared error**: The expected value of the square of the error between the fitted value (for the submodel) and the true conditional mean at \( x_i \):

\[ \text{MSE} (\hat{y}_i) = E([\hat{y}_i - E(Y \mid x_i)]^2). \]

**Note:**
1. \( \text{MSE} (\hat{y}_i) \) is defined like the sampling variance of \( \hat{y}_i \).
2. Thus, if \( \hat{y}_i \) is an unbiased estimator of \( E(Y \mid x_i) \), then \( \text{MSE} (\hat{y}_i) = \) _____________
3. Do not confuse with another use of MSE -- to denote \( \text{RSS}/\text{df} = \text{Mean Square for Residuals} \) (on regression ANOVA table)
4. MSE is *not* a statistic -- it involves the parameter \( E(Y \mid x_i) \).
We would like MSE ($\hat{y}_i$) to be small. To understand MSE better, we will examine, for fixed $i$, the variance of $\hat{y}_i - E(Y \mid x_i)$:

$$
\begin{align*}
\text{Var}(\hat{y}_i - E(Y \mid x_i)) &= E((\hat{y}_i - E(Y \mid x_i))^2) - [E(\hat{y}_i - E(Y \mid x_i))]^2 \\
 &= \text{MSE}(\hat{y}_i) - [E(\hat{y}_i) - E(Y \mid x_i)]^2 \\
 &= \text{MSE}(\hat{y}_i) - [\text{bias}(\hat{y}_i)]^2.
\end{align*}
$$

Also, since $E(Y \mid x_i)$ is constant,

$$
\text{Var}(\hat{y}_i - E(Y \mid x_i)) = \text{Var}(\hat{y}_i).
$$

Thus,

$$
\text{MSE}(\hat{y}_i) = \text{Var}(\hat{y}_i) + [\text{bias}(\hat{y}_i)]^2.
$$

So MSE really is a combined measure of variance and bias.

Summarizing: Deleting a term typically decreases $\text{Var}(\hat{y}_i)$ but increases bias. So we want to play these effects off against each other by minimizing MSE ($\hat{y}_i$). But we need to do this minimization for all $i$’s, so we consider the total mean squared error

$$
J = \sum_{i=1}^{n} \text{MSE}(\hat{y}_i)
$$

$$
= \sum_{i=1}^{n} \{\text{Var}(\hat{y}_i) + [\text{bias}(\hat{y}_i)]^2\}.
$$

We want this to be small. Since $J$ involves the parameters $E(Y \mid x_i)$, we need to estimate it. It works better to estimate the total normed mean squared error

$$
\gamma \text{ (or } \Gamma) = J/\sigma^2
$$

(whence $\sigma^2$ is as usual the conditional variance of the full model). Remember that $\hat{y}_i$ is the fitted value for the submodel, so $\gamma$ depends on the submodel. To emphasize this, we will denote $\gamma$ by $\gamma_I$, where $I$ is the set of terms retained in the submodel.

If the submodel is unbiased, then

$$
\gamma_I = (1/\sigma^2) \sum_{i=1}^{n} \text{Var}(\hat{y}_i),
$$

Now appropriate calculations show that
$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \text{Var}(\hat{y}_i) = k_i.$$  

(***)

the number of terms in I, whether or not the submodel is unbiased. (Try doing the calculation for $k_i = 2$ -- i.e., when the submodel is a simple linear regression model, using the formula for $\text{Var}(\hat{y}_i)$ in that case.) This implies that an unbiased model has $\gamma_i = k_i$

Thus having $\gamma_i$ close to $k_i$ implies that the submodel has small bias.

Summarizing: A good submodel has $\gamma_i$

(i) small (to get small total error)
(ii) near $k_i$ (to get small bias).

Putting together (**), (**), and (***) gives

$$\gamma_i = k_i + \frac{1}{\sigma^2} \sum_{i=1}^{n} [\text{bias (} \hat{y}_i \text{)]^2}.$$ 

It turns out that $(n - k_i)(\hat{\sigma}_i^2 - \hat{\sigma}_2^2)$ (where $\hat{\sigma}_i^2$ is the estimated conditional variance for the submodel) is an appropriate estimator for $\sum_{i=1}^{n} [\text{bias (} \hat{y}_i \text{)]^2}$, so the statistic

$$C_i = k_i + \frac{(n - k_i)(\hat{\sigma}_i^2 - \hat{\sigma}_2^2)}{\hat{\sigma}_2^2}$$

is an estimator of $\gamma_i$. $C_i$ is called Mallow's $C_i$ statistic. (It is sometimes called $C_p$, where $p = k_i$.) Some algebraic manipulation results in the alternate formulation

$$C_i = k_i + (n - k_i) \frac{\hat{\sigma}_i^2}{\hat{\sigma}_2^2} - (n - k_i)$$

$$= \frac{RSS_i}{\hat{\sigma}_2^2} + 2k_i - n.$$ 

Thus we can use Mallow's statistic to help identify good candidates for submodels by looking for submodels where $C_i$ is both

(i) small (suggesting small total error)
and
(ii) $\leq k_i$ (suggesting small bias)

Comments:

1. Mallow's statistic is provided by many software packages in some model-selection routine. Arc gives it in both Forward selection and Backward elimination. Other software
(e.g., Minitab) may use different procedures for Forward and Backward selection/elimination, but give Mallow's statistic in another routine (e.g., Best Subsets).

2. Since $C_i$ is a statistic, it will have sampling variability. It might happen, in particular, that $C_i$ is negative, which would suggest small bias. It also might happen that $C_i$ is larger than $k_i$ even when the model is unbiased, but there is no way to distinguish this situation from a case where there is bias but $C_i$ happens to be less than $\gamma_i$. 