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 $E(Y|\mathbf{x}) = \eta_0 + \eta_1 u_1 + \eta_2 u_2$. Then

$$Y = \eta_0 + \eta_1 u_1 + \eta_2 u_2 + \varepsilon. \text{ (So } \varepsilon \text{ is a random variable with } E(\varepsilon) = 0.)$$

Suppose further that

$$u_2 = 2u_1 + \delta, \text{ where } \delta \text{ is a random variable with } E(\delta) = 0.$$

Then

$$\begin{aligned} Y &= \eta_0 + \eta_1 u_1 + \eta_2(2u_1 + \delta) + \varepsilon \\ &= \eta_0 + (\eta_1 + 2\eta_2)u_1 + (\eta_2\delta + \varepsilon) \\ &= \eta_0' + \eta_1' u_1 + \varepsilon' \end{aligned}$$

where $\eta_0' = \eta_0$, $\eta_1' = \eta_1 + 2\eta_2$, and $\varepsilon' = \eta_2\delta + \varepsilon$. Since

$$E(\varepsilon') = E(\eta_2\delta + \varepsilon) = \eta_2 E(\delta) + E(\varepsilon) = 0,$$

the mean function for the submodel is

$$E(Y|\mathbf{x}) = \eta_0' + \eta_1' u_1.$$

Now suppose we fit both models by least squares, giving fits $\hat{y}_i$ for the full model and $\hat{y}_{i\text{sub}}$ 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 $\hat{y}_{i\text{sub}}$ are:

$E(\hat{y}_i) = \eta_0 + \eta_1 u_{i1} + \eta_2 u_{i2} = \eta_0 + \eta_1 u_{i1} + \eta_2(2u_{i1} + \delta_i)$ where δ_i is the value of δ for observation i, and

$$E(\hat{y}_{i\text{sub}}) = \eta_0' + \eta_1' u_{i1} = \eta_0 + (\eta_1 + 2\eta_2)\, u_{i1}.$$

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

Definition: The *bias* of an estimator is the difference between the expected value of the estimator and the parameter being estimated. So for parameter $E(Y \mid \mathbf{x}_i)$ and estimator $\hat{y}_{i\text{sub}}$,

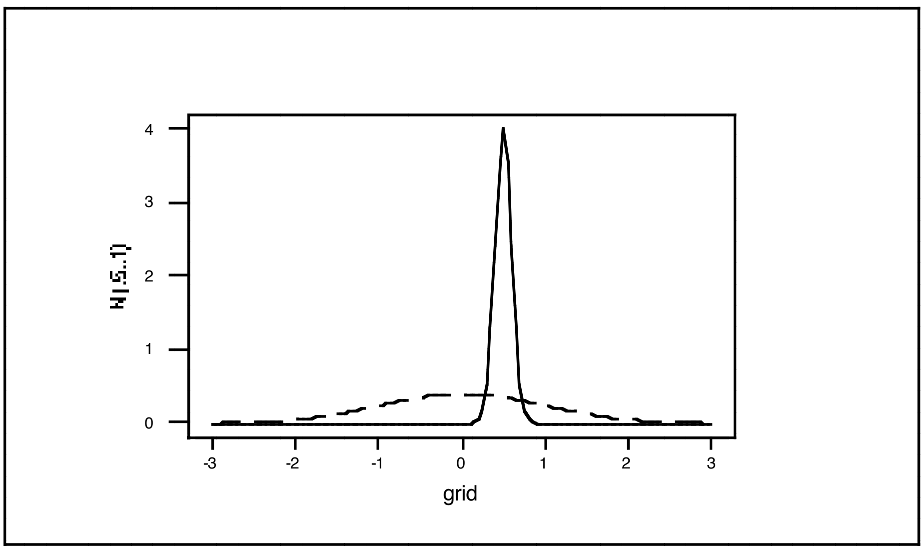
$$\text{bias}\,(\hat{y}_{i\text{sub}}) = E(\hat{y}_{i\text{sub}}) - E(Y \mid \mathbf{x}_i)$$

***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

$$\text{Var}(\hat{\eta}_j) = \frac{\sigma^2}{SU_jU_j}\frac{1}{1-R_j^2},$$

where SU_jU_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 Var($\hat{\eta}_j$); deleting a term usually decreases Var($\hat{\eta}_j$) (i.e., gives a more precise estimate of η_j). Since $\hat{y}_i$ is a linear combination of the $\hat{\eta}_j$'s, the effect will be the same for 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 $\mathbf{x}_i$:

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

Note:

1. 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 | $\mathbf{x}_i$), then MSE ($\hat{y}_i$) = _____________
3. Do not confuse with another use of MSE -- to denote RSS/df = Mean Square for Residuals (on regression ANOVA table)
4. MSE is *not* a statistic – it involves the parameter E(Y | $\mathbf{x}_i$).

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 | $\mathbf{x}_i$):

$$\begin{aligned}\mathrm{Var}(\hat{y}_i - E(Y \mid \mathbf{x}_i)) &= E([\hat{y}_i - E(Y \mid \mathbf{x}_i)]^2) - [E(\hat{y}_i - E(Y \mid \mathbf{x}_i))]^2 \\ &= \mathrm{MSE}(\hat{y}_i) - [E(\hat{y}_i) - E(Y \mid \mathbf{x}_i)]^2 \\ &= \mathrm{MSE}(\hat{y}_i) - [\mathrm{bias}\,(\hat{y}_i)]^2.\end{aligned}$$

Also, since E(Y | $\mathbf{x}_i$) is constant,

$$\mathrm{Var}(\hat{y}_i - E(Y \mid \mathbf{x}_i)) = \mathrm{Var}(\hat{y}_i).$$

Thus,

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

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

Summarizing: Deleting a term typically decreases 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*

$$\begin{aligned}J &= \sum_{i=1}^{n} \mathrm{MSE}\,(\hat{y}_i) \\ &= \sum_{i=1}^{n} \{\mathrm{Var}(\hat{y}_i) + [\mathrm{bias}\,(\hat{y}_i)]^2\}. \qquad (*)\end{aligned}$$

We want this to be small. Since J involves the parameters E(Y | $\mathbf{x}_i$), we need to estimate it. It works better to estimate the *total normed mean squared error*

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

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

If the submodel is unbiased , then

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

Now appropriate calculations show that

$$(1/\sigma^2)\sum_{i=1}^{n} \mathrm{Var}(\hat{y}_i) = k_I, \qquad (***)$$

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 Var($\hat{y}_i$) in that case.) This implies that an unbiased model has $\gamma_I = k_I$ Thus having γ_I close to k_I implies that the submodel has small bias.

Summarizing: A good submodel has γ_I

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

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

$$\gamma_I = k_I + (1/\sigma^2) \sum_{i=1}^{n} [\mathrm{bias}\ (\hat{y}_i)]^2.$$

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

$$C_I = k_I + \frac{(n - k_I)(\hat{\sigma}_I^2 - \hat{\sigma}^2)}{\hat{\sigma}^2}$$

is an estimator of γ_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

$$C_I = k_I + (n - k_I)\frac{\hat{\sigma}_I^2}{\hat{\sigma}^2} - (n - k_I)$$

$$= \frac{RSS_I}{\hat{\sigma}^2} + 2k_I - n.$$

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 γ_I.