M 374G/384G

SELECTING TERMS (Supplement to Section 11.5)

Consider a regression problem where $E(Y \mid \mathbf{x}) = \boldsymbol{\eta}^T\mathbf{u}$ is the correct model for the mean function. Often such a model has too many terms to be usable. Can some terms be deleted without important loss of information?

One problem that might result from dropping terms is that the resulting mean estimator might be biased. For example, if the *correct* model is $E(Y \mid \mathbf{x}) = \eta_0 + \eta_1 u_1 + \eta_2 u_2 + \ldots \eta_{k-1}u_{k-1}$ where $\eta_{k-1} \neq 0$ and we fit the model $E(Y \mid \mathbf{x}) = \gamma_0 + \gamma_1 u_1 + \gamma_2 u_2 + \ldots \gamma_{k-2}u_{k-2}$ by least squares to get fitted values $\hat{y}_i$, then (since the least squares estimates are unbiased for the model used),

$$E(\hat{y}_i) = \gamma_0 + \gamma_1 u_{i1} + \gamma_2 u_{i2} + \ldots \gamma_{k-2}u_{i,k-2},$$

which might *not* be the same as

$$\eta_0 + \eta_1 u_{i1} + \eta_2 u_{i2} + \ldots \eta_{k-1}u_{i,k-1} = E(Y \mid \mathbf{x}_i).$$

The difference between the expected value of the estimate and the parameter being estimated is called the *bias* of the estimator:

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

(Here, $\hat{y}_i$ is the estimate from the *sub*model.)

However, dropping terms might also reduce the variance. Sometimes, having biased estimates is the lesser of two evils. (Try drawing a picture to illustrate this.) 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 $\mathbf{x}_i$:

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

Note:

1. MSE $(\hat{y}_i)$ is deefined like the sampling variance of $\hat{y}_i$.
2. Thus, if $\hat{y}_i$ is an unbiased estimator of $E(Y \mid \mathbf{x}_i)$, 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)

We would like MSE $(\hat{y}_i)$ to be small. To understand MSE better, we will examine, for fixed i, the variance of $\hat{y}_i - E(Y \mid \mathbf{x}_i)$:

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

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

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

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

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. Now (see Section 10.1.5) the sampling variance of $\hat{\eta}_j$ in the submodel 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 $\hat{\eta}_j$). 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: 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*

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

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

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*

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

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

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 conditional variance of the submodel) is an appropriate estimator for $\sum_{i=1}^{n} [\mathrm{bias}\,(\hat{y}_i)]^2$, 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 selcetion 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.

2. Since C_I is a statistic, it will have sampling variability. It might happen, for example, 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.