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

$$Y = \eta_0 + \eta_1 u_1 + \eta_2 u_2 + \varepsilon.$$

(So ε is a random variable with E(ε) = 0.)

Suppose further that

$$u_2 = 2u_1 + \delta,$$

where δ is a random variable with E(δ) = 0.

Then

$$Y = \eta_0 + \eta_1 u_1 + \eta_2(2u_1 + \delta) + \varepsilon$$

$$= \eta_0 + (\eta_1 + 2\eta_2)u_1 + (\eta_2\delta + \varepsilon)$$

$$= \eta_0' + \eta_1' u_1 + \varepsilon'$$

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 $\hat{y}_{i\,\mathrm{sub}}$ 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 $\hat{y}_{i\,\mathrm{sub}}$ 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

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

Note that $E(\hat{y}_i)$ has the additional term $\eta_2\delta_i$ that $E(\hat{y}_{i\,\mathrm{sub}})$ doesn't have.

Thus, if the full model is the true model, then $\hat{y}_{i\,\mathrm{sub}}$ 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 $\hat{y}_{i\,\mathrm{sub}}$,

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

***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

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

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.

Note:

- 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:

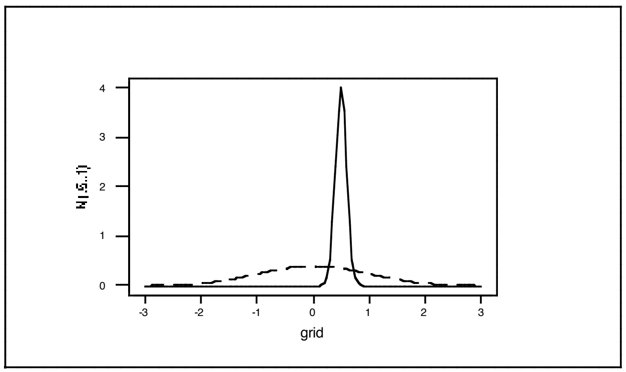
- Adding a term usually increases $\mathrm{Var}(\hat{\eta}_j)$
- Deleting a term usually decreases $\mathrm{Var}(\hat{\eta}_j)$ (i.e., gives a more precise estimate of η_j)
- Since $\hat{y}_i$ is a linear combination of the $\hat{\eta}_j$'s, the effect will be the same for $\mathrm{Var}(\hat{y}_i)$.

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 $\mathbf{x}_i$:

Definition: The *mean squared error* of a fitted value is

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

We'd like this to be small.

Comments:

1. MSE ($\hat{y}_i$) is defined like the sampling variance of $\hat{y}_i$, but using E(Y | $\mathbf{x}_i$) instead of E($\hat{y}_i$).

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 the MSE with earlier use of "Mean Squared Error" to mean Residual Mean Square (RSS/df).

4. MSE is not a statistic, since it involves the parameter E(Y | $\mathbf{x}_i$). We will eventually need to estimate it.

Details on MSE:

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

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

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

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

2) Since $E(Y \mid \mathbf{x}_i)$ 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 is a combined measure of variance and bias*.

Summarizing:

- Deleting a term typically decreases $\mathrm{Var}(\hat{y}_i)$ 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*

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

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

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

We want small J.

Note: If the submodel is unbiased, then each $\hat{y}_i$ will be unbiased, so J will = $\sum_{i=1}^{n} \mathrm{Var}(\hat{y}_i)$.

Since J involved the parameters $E(Y \mid \mathbf{x}_i)$, we need to estimate it.

It works better to estimate the *total normed mean squared error*

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

(σ^2 = the conditional variance of the *full* model).

Recall: $\hat{y}_i$ is the fitted value for the *submodel*.

Thus γ depends on the *submodel*.

Hence we call it γ_I, where I is the set of terms retained in the submodel.

If the submodel is *unbiased* , then

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

Appropriate calculations give

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

= number of terms in I.

(True for both biased and unbiased submodels)

This implies: In an unbiased model, $\gamma_I = k_I$

Thus: 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).

Combining (*), (**), and (***) gives

$$\gamma_I = J/\sigma^2$$

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

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

To estimate $\sum_{i=1}^{n} [\text{bias}(\hat{y}_i)]^2$, we can use

$$(n - k_I)(\hat{\sigma}_I^2 - \hat{\sigma}^2),$$

where $\hat{\sigma}_I^2$ = the estimated conditional variance for the submodel

Thus *Mallow's* C_I *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.

(It is sometimes called C_p, where $p = k_I$.)

Algebraic manipulation gives an alternate form:

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

$$= \frac{RSS_I}{\hat{\sigma}^2} + 2k_I - n. \quad (RSS_I = RSS_{sub})$$

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.

- C_I could be negative, suggesting small bias.

- C_I might be $> k_I$ even with an unbiased model, but we can't distinguish this from a case where there is bias but C_I happens to be less than γ_I.