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 \( \varepsilon \) is a random variable with \( E(\varepsilon) = 0 \).)

Suppose further that

\[ u_2 = 2u_1 + \delta, \]

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

Then

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

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

\[ = \eta_0' + \eta_1'u_1 + \varepsilon' \]

where \( \eta_0' = \eta_0, \eta_1' = \eta_1 + 2\eta_2, \) and \( \varepsilon' = \eta_2 \delta + \varepsilon \).

Since

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

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:

\[
E(\hat{Y}_i) = \eta_0 + \eta_1u_{i1} + \eta_2u_{i2} = \eta_0 + \eta_1u_{i1} + \eta_2(2u_{i1} + \delta_i)
\]

where \( \delta_i \) is the value of \( \delta \) 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 \( E(Y | x_i) \)

**Definition:** The bias of an estimator is the difference
between the expected value of the estimator and the
parameter being estimated.

So for parameter \( E(Y | x_i) \) and estimator \( \hat{Y}_{i\text{sub}} \),

\[
\text{bias} (\hat{Y}_{i\text{sub}}) = E(\hat{Y}_{i\text{sub}}) - E(Y | x_i).
\]

**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

\[
\text{Var}(\hat{\eta}_j) = \frac{\sigma^2}{SUU_j} \frac{1}{1 - R_j^2},
\]

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 \( \text{Var}(\hat{\eta}_j) \)
* 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 $x_i$:

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

$$\text{MSE} \left( \hat{y}_i \right) = E \left( (\hat{y}_i - E(Y \mid x_i))^2 \right).$$

We'd like this to be small.

**Comments:**

1. MSE $\left( \hat{y}_i \right)$ is defined like the sampling variance of $\hat{y}_i$, but using $E(Y \mid x_i)$ instead of $E(\hat{y}_i)$.

2. Thus, if $\hat{y}_i$ is an unbiased estimator of $E(Y \mid x_i)$, then

$$\text{MSE} \left( \hat{y}_i \right) =$$

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 $E(Y \mid x_i)$. We will eventually need to estimate it.
Details on MSE:

1) \[
\text{Var}(\hat{Y}_i - \text{E}(Y \mid x_i)) = E[(\hat{Y}_i - \text{E}(Y \mid x_i))^2] - [E(\hat{Y}_i - \text{E}(Y \mid x_i))]^2
\]
\[
= \text{MSE}(\hat{Y}_i) - [E(\hat{Y}_i) - \text{E}(Y \mid x_i)]^2
\]
\[
= \text{MSE}(\hat{Y}_i) - [\text{bias } (\hat{Y}_i)]^2.
\]

2) Since \( \text{E}(Y \mid x_i) \) is constant,
\[
\text{Var}(\hat{Y}_i - \text{E}(Y \mid x_i)) = \text{Var}(\hat{Y}_i).
\]

Thus,
\[
\text{MSE}(\hat{Y}_i) = \text{Var}(\hat{Y}_i) + [\text{bias } (\hat{Y}_i)]^2.
\]

So **MSE 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 \( \text{MSE}(\hat{Y}_i) \).

But we need to do this minimization for all \( i \)'s, so we consider the total mean squared error

\[
J = \sum_{i=1}^{n} \text{MSE}(\hat{Y}_i)
\]
\[
= \sum_{i=1}^{n} \{\text{Var}(\hat{Y}_i) + [\text{bias } (\hat{Y}_i)]^2\} \quad (*)
\]
\[
= \sum_{i=1}^{n} \text{Var}(\hat{Y}_i) + \sum_{i=1}^{n} [\text{bias } (\hat{Y}_i)]^2.
\]

We want small \( J \).

Note: If the submodel is unbiased, then each \( \hat{Y}_i \) will be unbiased, so \( J \) will = \( \sum_{i=1}^{n} \text{Var}(\hat{Y}_i) \).
Since $J$ involved the parameters $E(Y | x_i)$, we need to estimate it.

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

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

($\sigma^2$ = the conditional variance of the *full model*).

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

Thus $\gamma$ depends on the *submodel*.

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

If the submodel is *unbiased*, then

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

Appropriate calculations give

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

= number of terms in $I$.

(True for both biased and unbiased submodels)

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

Thus: $\gamma_I$ 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).
Combining (*) , (**), and (***) gives

\[ \gamma_1 = \frac{I}{\sigma^2} \]

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

\[ = k_1 + \frac{1}{\sigma^2} \sum_{i=1}^{n} [\text{bias}(\hat{y}_i)]^2. \]

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

\[ (n - k_1)(\hat{\sigma}^2 - \tilde{\sigma}^2), \]

where \( \hat{\sigma}^2 \) = the estimated conditional variance for the submodel

Thus Mallow’s \( C_1 \) statistic

\[ C_1 = k_1 + \frac{(n - k_1)(\hat{\sigma}^2 - \tilde{\sigma}^2)}{\hat{\sigma}^2} \]

is an estimator of \( \gamma_1. \)

(It is sometimes called \( C_p \), where \( p = k_1. \))

Algebraic manipulation gives an alternate form:

\[ C_1 = k_1 + (n - k_1) \frac{\hat{\sigma}_I^2}{\hat{\sigma}^2} - (n - k_1) \]

\[ = \frac{\text{RSS}_I}{\hat{\sigma}^2} + 2k_1 - n. \quad (\text{RSS}_I = \text{RSS}_{\text{sub}}) \]

Thus: We can use Mallow’s statistic to help identify good candidates for submodels by looking for submodels where \( C_1 \) is both

(i) small (suggesting small total error)

and

(ii) \( \leq k_1 \) (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 $\gamma_i$. 