TESTING FULL MODELS AGAINST SUBMODELS

(ref: Sections 11.2 – 11.2)

As in simple linear regression, we may want to test submodels against full models.

Example: Haystacks data. The model

$$E(\text{Vol}|C, \text{Over}) = \eta_0 + \eta_1 C^3 + \eta_2 \text{Over}^3$$

is a submodel of the larger cubic model

$$E(\text{Vol}|C, \text{Over}) = \eta_0 + \eta_1 C^3 + \eta_2 \text{Over}^3 + \eta_3 C^2\text{Over} + \eta_4 C\text{Over}^2$$

More generally: We may wish to test a submodel

$$E(Y|\underline{x}) = \eta_0 + \eta_1 u_1 + \ldots + \eta_l u_l$$

against a full model

$$E(Y|\underline{x}) = \eta_0 + \eta_1 u_1 + \ldots + \eta_{k-1} u_{k-1} \qquad (l \leq k-1).$$

Corresponding hypothesis test on coefficients:

NH:
AH:

Note:

- By re-ordering terms, this covers any situation where the null hypothesis is of the form "a certain set of coefficients is 0".
- Other types of tests of submodels can be handled, as in simple linear regression; we'll just discuss tests of this type.

Assuming:

- All four regression assumptions hold *for the model with all terms* **and**
- All four regression assumptions hold *with the desired terms omitted,*

then the test statistic is the same as in simple linear regression:

$$F = \frac{(RSS_{sub} - RSS_{full})/(df_{sub} - df_{full})}{\hat{\sigma}_{full}^{\,2}}$$

$$= \frac{(RSS_{sub} - RSS_{full})/(df_{sub} - df_{full})}{RSS_{full}/df_{full}}$$

$$= \frac{RSS_{sub} - RSS_{full}}{RSS_{full}} \bullet \frac{df_{full}}{df_{sub} - df_{full}}$$

$$\sim F(df_{sub} - df_{full}, df_{full}).$$

Recall: It is possible that the full model with all terms is linear, but when some terms are omitted, the conditional mean function might not be linear.

Example: True full model

$$E(Y|x_1, x_2) = 1 + 2x_1 + 3x_2.$$

Calculations similar to ones done earlier show

$$E(Y|x_1) = E(E(Y|\, x_1, x_2)|x_1)$$

$$= E(1 + 2x_1 + 3x_2|x_1)$$

$$= 1 + 2x_1 + 3E(x_2|x_1)$$

If, say, $E(x_2|x_1) = \log(x_1)$, then

$$E(Y|x_1) = 1 + 2x_1 + 3\log(x_1),$$

which is *not* linear in x_1.

Consequence: You cannot be confident of the results of an F-test if you have no reason to believe that you will still have a linear mean function after dropping the terms in question. *Be cautious*!

Note: It is also possible to invalidate the constant variance assumption by dropping terms; see Section 11.1.2, p. 265.

Unfortunately, many people don't realize that the model assumptions may be violated when dropping terms, so the F test is often applied when the conditions for it to be valid do not apply.

Moral: Be cautious when reading the literature.

However: Recall that *if* $U_1, U_2, \ldots, U_{k-1}$, Y are multivariate normal, then every marginal and conditional distribution is also multivariate normal, so the above problems will *not* occur in this case.

Moreover: The F-tests for submodels are fairly robust to departures from the linearity assumptions under *either* of the following conditions:

(i) The *terms are "linearly related*", i.e., $E(U_i|U_j)$ is a linear function of U_j for each pair i,j (and the other assumptions hold).

or

(ii) $U_1, U_2, \ldots, U_{k-1}$, Y are close to multivariate normal (and the other assumptions hold).

Practical Consequence: If you plan to consider submodels (common when dealing with many terms), then you should transform variables before using least squares and testing submodels. Try to get:

- Multivariate normality
- Or close to multivariate normality
- Or at least terms linearly related as much as possible.

Arc software can attempt to do this!

Comment: "Linearly related" includes the case of independent variables – e.g., if x_1 and x_2 are independent, then $E(x_1|x_2) = E(x_1) = \mu_1$ *is* a linear function of x_1.