# TESTING FULL MODELS AGAINST SUBMODELS (Ref: Sections 11.1-11.2)

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

*Example*: With the Haystacks data, we may consider the model

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

as a submodel of the larger cubic model

$$E(\mathrm{Vol}|C, \mathrm{Over}) = \eta_0 + \eta_1 C^3 + \eta_2 \mathrm{Over}^3 + \eta_3 C^2 \mathrm{Over} + \eta 4\mathrm{COver}2.$$

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

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

This corresponds to a hypothesis test on coefficients with

NH:
AH:

(Note that by rearranging 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, but we will just discuss tests of this type.)

*If* all regression assumptions hold for *both* the full model (all terms) *and* the submodel (certain terms omitted), the test statistic is the same as in simple linear regression:

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

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

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

As we have seen, it is possible for the full model with all terms to be linear, but that does not guarantee that when some terms are omitted, a linear model still fits.

*Example*: Suppose the true full model is

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

Then calculations similar to ones done earlier show

$$\begin{aligned} E(Y|x_1) &= E(E(Y|\ x_1, x_2)|x_1) \\ &= E(1 + 2x_1 + 3x_2|x_1) \\ &= 1 + 2x_1 + 3E(x_2|x_1) \end{aligned}$$

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 problem 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 (which is 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 $x_1$.