SUBMODELS (NESTED MODELS) AND ANALYSIS OF VARIANCE OF REGRESSION MODELS

We will assume we have data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and make the usual assumptions of independence and normality.

Our full model: (3 parameters)

$E(Y|x) = \eta_0 + \eta_1 x$

$Var(Y|x) = \sigma^2$

We have discussed how to "fit" the full model from data using least squares. We can also fit a submodel by least squares.

Example 1: To fit the submodel $E(Y|x) = 2 + \eta_1 x$

$Var(Y|x) = \sigma^2$,

consider lines $y = 2 + h_1 x$ and minimize

$RSS(h_1) = \sum d_i^2 = \sum [y_i - (2 + h_1 x_i)]^2$

to get η_1. [Draw a picture.]

Note: For this example, $y_i - (2 + h_1 x_i) = (y_i - 2) - h_1 x_i$,

so fitting this model is equivalent to fitting the model

$E(Y|x) = \eta_1 x$

$Var(Y|x) = \sigma^2$

to the transformed data $(x_1, y_1 - 2), (x_2, y_2 - 2), \dots, (x_n, y_n - 2)$

Example 2: For the submodel $E(Y|x) = \eta_0$

$Var(Y|x) = \sigma^2$,

we minimize $RSS(h_0) = \sum d_i^2 = \sum (y_i - h_0)^2$ [Draw a picture.]

- Carry out details
- Result: $h_0 = \bar{y}$ -- the same as the univariate estimate.
- Show that this is also the same as setting $\hat{\eta}_1 = 0$ in the least squares fit for the full model.

Caution: This phenomenon does *not* always happen, as the exercise below shows.

Exercise: Try finding the least squares fit for the submodel

$E(Y|x) = \eta_1 x$ ("Regression through the origin")

$Var(Y|x) = \sigma^2$

You should get a different formula for $\hat{\eta}_1$ that obtained by setting $\hat{\eta}_0$ = 0 in the formula for the least squares fit for the full model.

Generalizing: If we fit a submodel by Least Squares, we can define the residual sum of squares for the *submodel*:

$$RSS_{sub} = \sum(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i = \hat{E}_{sub}(Y|x)$ is the fitted value for the submodel.

Example: For the submodel in Example 2, $\hat{y}_i = \bar{y}$ for each i, so

$$RSS_{sub} = \sum(y_i - \bar{y})^2 = SYY$$

General Properties: (Stated without proof; true for multiple regression as well as simple regression)

- RSS_{sub} is a multiple of a χ^2 distribution, with
- degrees of freedom df_{sub} = n - (# of terms estimated), and
- $\hat{\sigma}_{sub}{}^2 = \dfrac{RSS_{sub}}{df_{sub}}$ is an estimate of σ^2 for the submodel.

Thus we can do infeerence tests using a submodel rather than the full model.

Another Perspective:

Example: The submodel $E(Y|x) = \eta_0$
$Var(Y|x) = \sigma^2$

Testing this model against the full model is equivalent to performing a hypothesis test with

NH: $\eta_1 = 0$
AH: $\eta_1 \neq 0$.

This hypothesis test uses the t-statistic

$$t = \frac{\hat{\eta}_1}{s.e.(\hat{\eta}_1)} = \frac{SXY/SXX}{\hat{\sigma}/\sqrt{SXX}} \sim t(n-2),$$

where here $\hat{\sigma} = \hat{\sigma}_{full}$ is the estimate of σ for the *full* model. Note that

$$t^2 = \frac{(SXY)^2 \big/ (SXX)^2}{\hat{\sigma}^2 \big/ SXX} = \frac{(SXY)^2}{\hat{\sigma}^2 (SXX)}$$

Recall:

$$RSS = SYY - \frac{(SXY)^2}{SXX}$$

$$RSS = RSS_{full}$$

$$SYY = RSS_{sub}$$

Thus

$$RSS_{sub} - RSS_{full} = \frac{(SXY)^2}{SXX}.$$

so

$$t^2 = \frac{RSS_{sub} - RSS_{full}}{\hat{\sigma}^2}$$

F Distributions

Recall: A t(k) random variable has the distribution of a random variable of the form

where

Thus

t^2 ~

Also,

Z^2 ~

Definition: An *F-distribution $F(\nu_1, \nu_2)$ with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator* is the distribution of a random variable of the form

$$\frac{W/\nu_1}{U/\nu_2}$$ where $W \sim \chi^2(\nu_1)$

$U \sim \chi^2(\nu_2)$

and U and W are independent.

Thus:

$$\frac{RSS_{sub} - RSS_{full}}{\hat{\sigma}^2} \sim F(1, n\text{-}2),$$

so we could also do our hypothesis test with an F-test.

Example: Forbes data.

Another way to look at the F-statistic:

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

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

i.e., F is the ratio of (the residual sum of squares for the submodel compared with the full model) and (the residual sum of squares for the full model) - - *but* with each divided by its degrees of freedom to "weight" them appropriately to get a tractable distribution.

More generally: Whenever we have a submodel (in multiple linear regression as well as simple linear regression),

a. RSS_{sub} (hence $\hat{\sigma}^2_{sub}$) will be a constant times a χ^2 distribution, with degrees of freedom df_{sub}, which we then also refer to as the degrees of freedom of RSS_{sub} and of $\hat{\sigma}^2_{sub}$.

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

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

Thus we can use an F statistic for the hypothesis test

NH: Submodel

AH: Full model