Nested Models

$$M_0: \quad y = \underbrace{\beta_0 + \beta_1 x_1 + ... + \beta_{q-1} x_{q-1}}_{\dim(M_0)=q} + \varepsilon$$

$$M_A: \quad y = \underbrace{\beta_0 + \beta_1 x_1 + ... + \beta_{q-1} x_{q-1} + \beta_q x_q + ... + \beta_{p-1} x_{p-1}}_{\dim(M_A)=p} + \varepsilon$$

<u>want to test the following null hypothesis</u>

$$H_0: \quad \beta_q = ... = \beta_{p-1} = 0$$

The F-Test

Theory Under M_0 (and normality of all y_i),

$$\frac{(\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2 - \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2)/(p-q)}{\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2/(n-p)} \sim F_{(p-q,n-p)}.$$

So reject M_0 in favor of M_A if LHS is large, when measured by an F-distribution.

Ockham's Razor

Latin: Pluralitas non est ponenda sine necesitate.

—William of Ockham (1285–1349)

English: Pluralities should not be posited without necessity.

Just A Little More Detail

Numerator By Ockham's razor, "makes sense" to focus on the difference

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2 - \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2.$$

Denominator Needs to be "orthogonal" to numerator for F-distribution, but projection geometry "clearly" shows

$$\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_0\|^2 - \|\boldsymbol{y} - \widehat{\boldsymbol{y}}_A\|^2 = \|\widehat{\boldsymbol{y}}_0 - \widehat{\boldsymbol{y}}_A\|^2$$

and

$$\widehat{\boldsymbol{y}}_0 - \widehat{\boldsymbol{y}}_A \perp \boldsymbol{y} - \widehat{\boldsymbol{y}}_A.$$

Special Case: $p = q + 1$

$$M_0: \quad y = \beta_0 + \beta_1 x_1 + ... + \beta_{q-1} x_{q-1} + \varepsilon$$

$$M_A: \quad y = \beta_0 + \beta_1 x_1 + ... + \beta_{q-1} x_{q-1} + \beta_q x_q + \varepsilon$$

T vs F

Expect t-test of β_q=0 to be equivalent to F-test.

Special Case: $p = q + 1$

$$M_0: \quad y = \beta_0 + \beta_1 x_1 + ... + \beta_{q-1} x_{q-1} + \varepsilon$$

$$M_A: \quad y = \beta_0 + \beta_1 x_1 + ... + \beta_{q-1} x_{q-1} + \beta_q x_q + \varepsilon$$

T vs F

Expect t-test of β_q=0 to be equivalent to F-test.

Simple Example

Consider the set of nested models below.

$$M_0: y = \alpha + \varepsilon \qquad \text{vs} \qquad M_A: y = \alpha + \beta x + \varepsilon$$

Let T_β be the t-statistic for testing $\beta = 0$ and F_β, the F-statistic for testing M_0 against M_A. Then, $T_\beta^2 = F_\beta$.

Details

$$M_A \Rightarrow \widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}, \quad \widehat{\beta} = \ldots \Rightarrow \widehat{y}_i^{(A)} = \widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$$

$$M_0 \Rightarrow \widehat{\alpha} = \bar{y}, \quad \beta = 0 \Rightarrow \widehat{y}_i^{(0)} = \bar{y}$$

$$T_\beta = \frac{\widehat{\beta}}{\sqrt{\dfrac{\widehat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \quad \text{whereas} \quad F_\beta = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2}{\dfrac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{n-2}}$$

$$\text{suffices if} \quad \sum_{i=1}^n (\widehat{y}_i - \bar{y})^2 = \widehat{\beta}^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\Downarrow \qquad \Uparrow$$

$$\text{indeed} \quad \sum_{i=1}^n (\underbrace{\widehat{\alpha} + \widehat{\beta}x_i}_{\widehat{y}_i} - \bar{y})^2 = \sum_{i=1}^n (\underbrace{\bar{y} - \widehat{\beta}\bar{x}}_{\widehat{\alpha}} + \widehat{\beta}x_i - \bar{y})^2$$

T- and F-Tests $\Leftrightarrow$ LRT

Example In the same spirit as Exercise 7.2 (p. 128), can show, for

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2),$$

(a)

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \left[\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right] + \widehat{\beta}^2 \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right],$$

(b)

$$2\log\Lambda(\beta) = n\log\left[1 + \frac{\big(T(\beta)\big)^2}{n-2}\right] \to \big(T(\beta)\big)^2 \text{ as } n \to \infty,$$

where $\Lambda(\beta)$ and $T(\beta)$ are respectively the LR- and t-statistics for testing $H_0 : \beta = 0$. [*<u>Remark</u>: If you try it, remember that the MLEs of α, σ^2 are different with and without the restriction $\beta = 0$.*]

K-Fold Cross Validation

1. randomly partition the data set into K groups, $\mathcal{G}_1, ..., \mathcal{G}_K$

2. **for** each $k = 1, 2, ..., K$

$$\widehat{\boldsymbol{\theta}}^{(-\mathcal{G}_k)} = \arg\min_{\boldsymbol{\theta}} \sum_{i \notin \mathcal{G}_k} [y_i - f(\boldsymbol{x}_i; \boldsymbol{\theta})]^2$$

$$\text{err}(k) = \sum_{i \in \mathcal{G}_k} \left[y_i - \widehat{y}_i^{(-\mathcal{G}_k)} \right]^2 = \sum_{i \in \mathcal{G}_k} \left[y_i - f\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}^{(-\mathcal{G}_k)} \right) \right]^2$$

   end for

3. assess the overall prediction error of model f as

$$\text{Err}(f) = \text{err}(1) + \text{err}(2) + ... + \text{err}(K)$$

Remark Do this for a number of candidate models and choose the one with smallest "Err." Usually implemented w/ $K = 2, 5, 10$.

Leave-One-Out CV

Special Case $K = n$ (aka n-fold CV)

$$\mathcal{G}_1 = \{1\}, \quad \mathcal{G}_2 = \{2\}, \quad ..., \quad \mathcal{G}_n = \{n\}$$

Theorem For $f(\boldsymbol{x};\boldsymbol{\theta})$ = linear regression model (in fact, any model such that $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$ for some $\boldsymbol{H}$ not depending on $\boldsymbol{y}$),

$$y_i - \widehat{y}_i^{(-i)} = \frac{y_i - \widehat{y}_i}{1 - \boldsymbol{H}_{ii}}.$$

Remark Can do leave-one-out (n-fold) CV without iteration.

Proof

$$
\begin{bmatrix} \cancel{\widehat{y}_1}\ \widehat{y}_1^{(-i)} \\ \vdots \\ \cancel{\widehat{y}_i}\ \widehat{y}_i^{(-i)} \\ \vdots \\ \cancel{\widehat{y}_n}\ \widehat{y}_n^{(-i)} \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{H}_{i1} & \cdots & \boldsymbol{H}_{ii} & \cdots & \boldsymbol{H}_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ \cancel{y_i}\ \widehat{y}_i^{(-i)} \\ \vdots \\ y_n \end{bmatrix}
$$

$$
\widehat{y}_i = \sum_{j=1}^{n} \boldsymbol{H}_{ij} y_j \quad \Rightarrow \quad \widehat{y}_i^{(-i)} = \sum_{j=1}^{n} \boldsymbol{H}_{ij} y_j - \boldsymbol{H}_{ii} y_i + \boldsymbol{H}_{ii} \widehat{y}_i^{(-i)}
$$

$$
y_i - \text{LHS} = y_i - \text{RHS}
$$

$$
\Rightarrow \quad y_i - \widehat{y}_i^{(-i)} = y_i - \widehat{y}_i + \boldsymbol{H}_{ii} (y_i - \widehat{y}_i^{(-i)})
$$

Generalized Cross Validation

$$\text{CV} = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i^{(-i)} \right)^2 = \sum_{i=1}^{n} \left(\frac{y_i - \widehat{y}_i}{1 - \boldsymbol{H}_{ii}} \right)^2$$

$$\text{"average } \boldsymbol{H}_{ii}\text{"} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}_{ii} = \frac{1}{n} \text{tr}(\boldsymbol{H}) = \frac{p}{n}$$

$$\text{GCV} = \sum_{i=1}^{n} \left(\frac{y_i - \widehat{y}_i}{1 - p/n} \right)^2$$

$$\Uparrow$$

a penalty on model size

Akaike Information Criterion (AIC)

Exercise If σ^2 is known, the so-called Akaike Information Criterion (AIC) for evaluating a linear regression model can be equivalently expressed by

$$\text{AIC} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + 2p\sigma^2,$$

which, again, puts a penalty on model size. Use the Taylor approximation, $1/(1-u)^2 \approx 1 + 2u$ for u small, to explain why GCV and AIC are "more or less" equivalent to each other when n is relatively large.

Remark Usually, one ends up with similar answers when choosing a model according to either CV, GCV, or AIC.