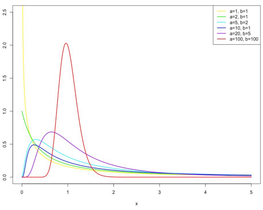
$\Gamma(n) = (n-1)!$ (if n is an integer)

$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x}\,dx$ (if n is not an integer)

The F-Distribution

$X \sim F_{(a,b)} \longrightarrow f(x) = \dfrac{\Gamma\left(\frac{a+b}{2}\right)}{\Gamma\left(\frac{a}{2}\right)\Gamma\left(\frac{b}{2}\right)} \left(\frac{a}{b}\right)^{a/2} x^{a/2-1} \left(1+\frac{a}{b}x\right)^{-\frac{a+b}{2}}, \quad x \geq 0$



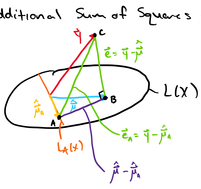
Where does this come from?

- Let $U \sim \chi^2_{(a)}$ and $V \sim \chi^2_{(b)}$
- Assume U and V are independent
- Then $X = \dfrac{U/a}{V/b} \sim F_{(a,b)}$

Some interesting properties:

- If $X \sim F_{(a,b)}$ then $\dfrac{1}{X} \sim F_{(b,a)}$
- If $Y \sim t_{(a)}$ then $Y^2 \sim F_{(1,a)}$ (*)

Returning to the Additional Sum of Squares principle....



Notice: $BC^2 = \vec{e}^T\vec{e} = SSE$ ← Sum of squared error in the full model

$AB^2 = \|\hat{\vec{y}} - \hat{\vec{y}}_A\|^2$

$AC^2 = \vec{e}_A^T\vec{e}_A \equiv SSE_A$ ← Sum of squared error in the reduced model

By Pythagorean Theorem $AC^2 = AB^2 + BC^2$

$SSE_A = \|\hat{\vec{y}} - \hat{\vec{y}}_A\|^2 + SSE$

$\Rightarrow \|\hat{\vec{y}} - \hat{\vec{y}}_A\|^2 = SSE_A - SSE$ * Additional sum of squares that the reduced model fails to account for

Since $\|\hat{\vec{y}} - \hat{\vec{y}}_A\|^2 \geq 0 \longrightarrow SSE_A \geq SSE$

SSE in the reduced model is always bigger than or equal to the SSE in the full model. Therefore the full model will always explain more response variation than the reduced model.

Using this development, the test statistic for the F-test can be rewritten as follows:

$$t = \frac{(SSE_A - SSE)/k}{SSE/(n-p-1)}$$

Example: Return to the "Sales" example

Suppose we want to test $H_0: \beta_1 = \beta_2 = 0$. This hypothesis implies that x_1 (promotional expenditures) and x_{12} (district potential score) do not significantly influence y (sales).

(a) This hypothesis can stated as

$H_0: A\vec{\beta} = \vec{0}$ vs. $H_A: A\vec{\beta} \neq \vec{0}$

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$k \times (p+1)$

(b) Calculate the test statistic

$SSE = 262$, $n-p-1 = 10$

$SSE_A = 535$, $k = 2$

$$t = \frac{(535-262)/2}{262/10} = 5.2$$

(c) Calculate the p-value

p-value $= P(T \geq 5.2)$ where $T \sim F_{(2,10)}$

$= 0.0283$

∴ we reject H_0 at a 5% significance level. And so we conclude that at least one of x_1 and/or x_{12} should be included in the model.

*Note 1: We can use this approach to test hypotheses about single βs (i.e. $H_0: \beta_j = 0$ vs. $H_A: \beta_j \neq 0$) and it turns out that the test is statistically equivalent to the t-test of the same hypothesis. This is due to property (*)

*Note 2: We can apply this approach to test the overall significance of the linear regression:

$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$ vs. $H_A: \beta_j \neq 0$ for some $j = 1, 2, \ldots, p$

If this null hypothesis is true, it implies that the response variable does not significantly depend on any of the chosen explanatory variable. If H_0 is not true then we can conclude that at least one of the explanatory variables is significantly influential.

Example: Sales

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs. $H_A: \beta_j \neq 0$ for some $j = 1, 2, 3, 4$

This is equivalent to

$H_0: A\vec{\beta} = \vec{0}$ vs $H_A: A\vec{\beta} \neq \vec{0}$

where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$SSE = 262$, $n-p-1 = 10$

$SSE_A = 89547$, $k = 4$

$$t = \frac{(89547-262)/4}{262/10} = 851.7198$$

p-value $= P(T \geq 851.7198)$ where $T \sim F_{(4,10)}$

$= 1.285 \times 10^{-12}$

∴ we reject H_0 and conclude that at least one of x_1, x_2, x_3, x_4 significantly influences y.