ANalysis Of VAriance (ANOVA)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\boldsymbol{0},\ \sigma^2\boldsymbol{I})$$

Let $\boldsymbol{X}_1 = \boldsymbol{1},\ \boldsymbol{X}_m = \boldsymbol{X},$ and $\boldsymbol{X}_{m+1} = \boldsymbol{I}$.

Suppose $\boldsymbol{X}_2,\ \ldots, \boldsymbol{X}_{m-1}$ are design matrices satisfying

$\mathcal{C}(\boldsymbol{X}_1) \subset \mathcal{C}(\boldsymbol{X}_2) \subset \cdots \subset \mathcal{C}(\boldsymbol{X}_{m-1}) \subset \mathcal{C}(\boldsymbol{X}_m)$.

Let $r_j = \text{rank}(\boldsymbol{X}_j)\ \forall\, j = 1, \ldots, m+1.$

Let $\boldsymbol{P}_j = \boldsymbol{P}_{\boldsymbol{X}_j} \quad \forall\, j = 1, \ldots, m+1$. Then

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y}.)^2 &= \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} = \boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_1)\boldsymbol{y} \\
&= \boldsymbol{y}'\left(\sum_{j=2}^{m+1} \boldsymbol{P}_j - \sum_{j=1}^{m} \boldsymbol{P}_j\right)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_m + \boldsymbol{P}_m - \boldsymbol{P}_{m-1} + \cdots + \boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_m)\boldsymbol{y} + \ldots + \boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y} \\
&= \sum_{j=1}^{m} \boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}.
\end{aligned}$$

The sums of squares in the equation

$$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}$$

are often arranged in an ANOVA table.

Sum of Squares	Sum of Squares
$\boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$SS(2 \mid 1)$
$\boldsymbol{y}'(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$	$SS(3 \mid 2)$
$\vdots$	$\vdots$
$\boldsymbol{y}'(\boldsymbol{P}_m - \boldsymbol{P}_{m-1})\boldsymbol{y}$	$SS(m \mid m-1)$
$\boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_m)\boldsymbol{y}$	$SSE = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}$
$\boldsymbol{y}'(I - \boldsymbol{P}_1)\boldsymbol{y}$	$SSTo = \sum_{i=1}^{n}(y_i - \bar{y}.)^2$

Note that $\forall\, j = 1, \ \ldots, \ m$

$$\begin{aligned}(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j) &= \boldsymbol{P}_{j+1}\boldsymbol{P}_{j+1} - \boldsymbol{P}_{j+1}\boldsymbol{P}_j - \boldsymbol{P}_j\boldsymbol{P}_{j+1} + \boldsymbol{P}_j\boldsymbol{P}_j \\ &= \boldsymbol{P}_{j+1} - \boldsymbol{P}_j - \boldsymbol{P}_j + \boldsymbol{P}_j \\ &= \boldsymbol{P}_{j+1} - \boldsymbol{P}_j.\end{aligned}$$

Also, $\forall\, j < \ell$

$$\begin{aligned}(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)(\boldsymbol{P}_{\ell+1} - \boldsymbol{P}_\ell) &= \boldsymbol{P}_{j+1}\boldsymbol{P}_{\ell+1} - \boldsymbol{P}_{j+1}\boldsymbol{P}_\ell - \boldsymbol{P}_j\boldsymbol{P}_{\ell+1} + \boldsymbol{P}_j\boldsymbol{P}_\ell \\ &= \boldsymbol{P}_{j+1} - \boldsymbol{P}_{j+1} - \boldsymbol{P}_j + \boldsymbol{P}_j \\ &= \boldsymbol{0}.\end{aligned}$$

Using these facts and previous facts about distributions of quadratic forms, it can be shown that

$$\frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}}{\sigma^2} \sim \chi^2_{r_{j+1}-r_j}(\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}/\sigma^2)$$

for all $j = 1, \ldots, m$ and that these m χ^2 random variables are mutually independent.

Sum of Squares	Degrees of Freedom	DF
$\boldsymbol{y}'(\boldsymbol{P}_2 - \boldsymbol{P}_1)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_2) - \text{rank}(\boldsymbol{X}_1)$	$r_2 - 1$
$\boldsymbol{y}'(\boldsymbol{P}_3 - \boldsymbol{P}_2)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_3) - \text{rank}(\boldsymbol{X}_2)$	$r_3 - r_2$
$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{y}'(\boldsymbol{P}_m - \boldsymbol{P}_{m-1})\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_m) - \text{rank}(\boldsymbol{X}_{m-1})$	$r - r_{m-1}$
$\boldsymbol{y}'(\boldsymbol{P}_{m+1} - \boldsymbol{P}_m)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_{m+1}) - \text{rank}(\boldsymbol{X}_m)$	$n - r$
$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y}$	$\text{rank}(\boldsymbol{X}_{m+1}) - \text{rank}(\boldsymbol{X}_1)$	$n - 1$

For $j = 1, \ldots, m-1$ we have

$$F_j = \frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}/(r_{j+1} - r_j)}{\boldsymbol{y}'(I - \boldsymbol{P}_X)\boldsymbol{y}/(n-r)}$$

$$\sim F_{r_{j+1}-r_j, n-r}(\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}/\sigma^2).$$

For $j = 1, \ldots, m-1$, define

$$MS(j+1 \mid j) = \frac{SS(j+1 \mid j)}{r_{j+1} - r_j} = \frac{\boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y}}{r_{j+1} - r_j}.$$

ANOVA Table

Sum of Squares	Degrees of Freedom	Mean Square
$SS(2 \mid 1)$	$r_2 - 1$	$MS(2\|1)$
$SS(3 \mid 2)$	$r_3 - r_2$	$MS(3\|2)$
$\vdots$	$\vdots$	$\vdots$
$SS(m \mid m-1)$	$r - r_{m-1}$	$MS(m\|m-1)$
SSE	$n - r$	MSE
$SSTO$	$n - 1$	

Note that

$$
\begin{aligned}
SS(j+1 \mid j) &= \boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j)\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{P}_{j+1} - \boldsymbol{P}_j + \boldsymbol{I} - \boldsymbol{I})\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_j - \boldsymbol{I} + \boldsymbol{P}_{j+1})\boldsymbol{y} \\
&= \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_j)\boldsymbol{y} - \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{j+1})\boldsymbol{y} \\
&= SSE_{\text{REDUCED}} - SSE_{\text{FULL}}
\end{aligned}
$$

Thus, $SS(j+1 \mid j)$ is the amount the error sum of square decreases when $\boldsymbol{y}$ is projected onto $\mathcal{C}(\boldsymbol{X}_{j+1})$ instead of $\mathcal{C}(\boldsymbol{X}_j)$.

$SS(j+1 \mid j),\ j = 1,\ \ldots,\ m-1$ are called *"Sequential Sums of Squares."*

SAS calls these "Type I Sums of Squares. "

The statistic

$$F_j = \frac{MS(j+1 \mid j)}{MSE}$$

can be used to test

$$H_0 : \mathrm{E}(\boldsymbol{y}) \in \mathcal{C}(\boldsymbol{X}_{\mathrm{j}}) \text{ vs. } \mathrm{H_A} : \mathrm{E}(\boldsymbol{y}) \in \mathcal{C}(\boldsymbol{X}_{\mathrm{j+1}}) \setminus \mathcal{C}(\boldsymbol{X}_{\mathrm{j}}).$$

The noncentrality parameter is

$$\begin{aligned}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}/\sigma^2 &= \frac{\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)'(\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2} \\ &= \mid\mid (\boldsymbol{P}_{j+1}-\boldsymbol{P}_j)\boldsymbol{X}\boldsymbol{\beta} \mid\mid^2 /\sigma^2 \\ &= \mid\mid \boldsymbol{P}_{j+1}\mathrm{E}(\boldsymbol{y}) - \boldsymbol{P}_{\mathrm{j}}\mathrm{E}(\boldsymbol{y}) \mid\mid^2 /\sigma^2.\end{aligned}$$

If H_0 is true, $\boldsymbol{P}_{j+1}\mathrm{E}(\boldsymbol{y}) = \boldsymbol{P}_{\mathrm{j}}\mathrm{E}(\boldsymbol{y}) = \mathrm{E}(\boldsymbol{y})$.

Thus, the $NCP = 0$ under H_0.

Example: Multiple Regression

$$
\begin{aligned}
\boldsymbol{X}_1 &= \mathbf{1} \\
\boldsymbol{X}_2 &= [\mathbf{1},\ \boldsymbol{x}_1] \\
\boldsymbol{X}_3 &= [\mathbf{1},\ \boldsymbol{x}_1,\ \boldsymbol{x}_2] \\
&\vdots \\
\boldsymbol{X}_m &= [\mathbf{1},\ \boldsymbol{x}_1,\ \ldots,\ \boldsymbol{x}_{m-1}]
\end{aligned}
$$

$SS(j+1 \mid j)$ is the decrease in SSE that results when the explanatory variable x_j is added to a model containing an intercept and explanatory variables $x_1, \ldots, x_{j-1}$.

Example: Test for Linear Trend and Test for Lack of Linear Fit

$$\boldsymbol{X}_1 = \mathbf{1}, \quad \boldsymbol{X}_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \\ 1 & 3 \end{bmatrix}, \quad \boldsymbol{X}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

F Test for a Linear Trend

Let $\mu_i =$ mean yield for a plot that received i units of fertilizer $(i = 1, 2, 3)$.

$$\frac{MS(2|1)}{MSE}$$

can be used to test

$$H_0 : \mu_1 = \mu_2 = \mu_3 \iff \mu_i = \beta_0 \; \forall \; i = 1, 2, 3 \text{ for some } \beta_0 \in \mathbb{R}$$

versus

$$H_A : \mu_i = \beta_0 + \beta_1(i) \quad i = 1, 2, 3 \text{ for some } \beta_0 \in \mathbb{R}, \; \beta_1 \in \mathbb{R} \setminus \{0\}.$$

F Test for Lack of Linear Fit

The statistic

$$\frac{MS(3|2)}{MSE}$$

can be used to test

$$H_0 : \mu_i = \beta_0 + \beta_1(i) \quad i = 1, 2, 3 \text{ for some } \beta_0, \beta_1 \in \mathbb{R}$$

versus

$$H_A : \text{There does not exist } \beta_0, \beta_1 \in \mathbb{R} \text{ such that}$$

$$\mu_i = \beta_0 + \beta_1(i) \qquad \forall\, i = 1, 2, 3.$$

The lack of fit test is a reduced vs. full model F test.

Thus, we can also obtain this test by testing

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} \quad \text{vs.} \quad H_A : \boldsymbol{C\beta} = \boldsymbol{d}$$

for appropriate $\boldsymbol{C}$ and $\boldsymbol{d}$.

$$\boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} \qquad \boldsymbol{C} = ? \qquad \boldsymbol{d} = ?$$