$\boxed{\text{ANOVA}}$ Summary

(Can be applied to $\boxed{\text{NESTED}}$ model selection)

→ Analysis Of Variance

one-way

[E.g.] IQ of different schools (單因素)

$\boxed{\text{Testing}}$: $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ H_A: H_0 is false

School 1	IQ_{11} $\cdots$	IQ_{1n_1}	n_1	$\overline{IQ}_1$
School 2	IQ_{21} $\cdots$	IQ_{2n_2}	n_2	$\overline{IQ}_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
School K	IQ_{K1}.	IQ_{Kn_K}	n_K	$\overline{IQ}_K$

$\overline{IQ}$

two-way

IQ of different schools & sex (双因素)

(Similar to 單因素 scenario)

$\left[\begin{array}{l}\text{school 1} \\ \vdots \\ \text{school K}\end{array}\right.$

$\left[\begin{array}{l}\text{Male} \\ \text{Female}\end{array}\right.$

(F-test)

① Understanding (One-way) from $\boxed{\text{LR perspective}}$

$\boxed{\text{Condition}}$

a) suppose G-M condition

$\begin{cases} ① \ E[\varepsilon_i] = 0 \\ ② \ cov(\varepsilon_i, \varepsilon_j) = 0 \end{cases}$

b) $\varepsilon \sim$ Gaussian

$\boxed{\text{Problem Setting:}}$ $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$

→ avoid col-linearity issue

$\boxed{X_{1i}, \ldots, X_{p-1,i} \rightsquigarrow 0/1}$

→ $\boxed{\text{dummy variable}}$

$\Leftrightarrow$ $\boxed{\text{group } 1 \; \vdots \; \text{group } p}$ - - - - -

- - - - -

since here we only have p-1 variables !

$\hat{Y}_i$ is derived from full model

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

$$\boxed{\text{LR Framework}} \Rightarrow \boxed{\frac{SSR / p-1}{SSE / n-p} \sim F}$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

Logic: ① we achieve SSR / SSE under full model

② $\frac{SSR / p}{SSE / n-p-1} \sim F$ under null assumption

→ $\boxed{\text{Q: Here, what is } \hat{Y}_i \text{ ?}}$

→ Answer: $\boxed{\text{ANOVA Test}}$ $\Leftrightarrow$ $\begin{cases} H_0: \beta_1 = \cdots = \beta_{p-1} = 0 \\ H_A: o/w \end{cases}$ $\Leftrightarrow$ $\frac{SSR / p-1}{SSE / n-p} \sim F$

→ ① Notice that, under H_0, model turns to

$$Y_i = \beta_0 + \varepsilon_i$$

$\Rightarrow \hat{\beta}_0 = \bar{Y}$

→ ② Try to figure out $\hat{Y}_i$!

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_{p-1} X_{p-1,i} + \varepsilon_i$$

$$\rightarrow \hat{\beta} = \min_{\beta} \| \hat{Y} - Y \|_2^2$$

$$= \min_{\beta} \sum_{Y \in G_1} (Y - \beta_0 - \beta_1)^2 + \cdots + \sum_{Y \in G_{p-1}} (Y - \beta_0 - \beta_{p-1})^2 + \sum_{Y \in G_p} (Y - \beta_0)^2$$

$$\Rightarrow \begin{cases} \hat{\beta}_0 = \bar{Y}_p \\ \hat{\beta}_0 + \hat{\beta}_1 = \bar{Y}_1 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_{p-1} = \bar{Y}_{p-1} \end{cases} \Rightarrow \boxed{\begin{array}{l} \hat{Y}_i = \bar{Y}_{g(i)} \\ g(i) = \{ \text{Group: } Y_i \text{ belongs to} \} \end{array}}$$

$\Rightarrow$ Therefore, we can compute $\begin{cases} SSR \\ SSE \end{cases}$

① $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

$= \sum_{k=1}^{p} (\#\text{ of data in Group } k) \cdot (\bar{y}_k - \bar{y})^2$

② $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$= \sum_{k=1}^{p} \sum_{y_i \in \text{Group } k} (y_i - \bar{y}_k)^2$

$\Rightarrow \dfrac{\dfrac{SSR}{p-1}}{\dfrac{SSE}{n-p}} \sim F(p-1, n-p)$

組间

組内

Another Perspective : Analyze Directly (Common Manner)

$\rightarrow$ From 学弱猹

Notation

$\begin{cases} \mu \rightarrow \text{mean of all group} \\ a_i \rightarrow i\text{-th group power} \end{cases}$

① Model Setting

$\begin{cases} y_{ij} = \mu + a_i + \varepsilon_{ij} & i = 1, 2, \ldots, r \; ; \; j = 1, 2, \ldots, m \\ \sum_{i=1}^{r} a_i = 0 \\ \mu_i = \mu + a_i \quad (y_{ij} = \mu_i + \varepsilon_{ij}) \\ \varepsilon_{ij} \sim N(0, \sigma^2) \; \& \text{ independent with each other} \end{cases}$

ANOVA $\Leftrightarrow$ Hypothesis Testing on $\begin{cases} H_0: a_1 = a_2 = \cdots = a_r = 0 \\ H_a: \text{o/w} \end{cases}$

② Analysis :

$$\begin{cases} \bar{Y}_i = \mu + a_i + \bar{\varepsilon}_i \\ \bar{Y} = \mu + \bar{\varepsilon} \end{cases}$$

$$SSR = \sum_{i=1}^{r} m \cdot (\bar{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{r} m (\bar{\varepsilon}_i + a_i - \bar{\varepsilon})^2$$

$$\overset{H_0}{=\!=} \sum_{i=1}^{r} m (\bar{\varepsilon}_i - \bar{\varepsilon})^2 \sim \sigma^2 \chi^2 (r-1)$$

$$SSE = \sum_{i=1}^{r} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_i)^2$$

$$= \sigma^2 \left(\chi_1^2 (m-1) + \cdots + \chi_r^2 (m-1) \right)$$

$$= \sigma^2 \chi^2 (r \cdot m - r)$$

Notice that $\begin{cases} (Y_{ij} - \bar{Y}_i)^2 \perp\!\!\!\perp \bar{Y}_i \\ (Y_{ij} - \bar{Y}_i)^2 \perp\!\!\!\perp \bar{Y}_\ell \quad \ell \neq i \end{cases} \Rightarrow (Y_{ij} - \bar{Y}_i)^2 \perp\!\!\!\perp (\bar{Y}_t - \bar{Y})^2 \quad \forall t$

$\Rightarrow$ $\underline{SSR \perp\!\!\!\perp SSE}$

$$\Rightarrow \frac{\frac{SSR}{r-1}}{\frac{SSE}{r(m-1)}} \sim F(r-1, rm-r)$$

$\boxed{\text{ANOVA Table}}$

	value	df	Normalize	F	P
SSR	$\sum_{i=1}^{r} m (\bar{Y}_i - \bar{Y})^2$	$r-1$	$SSR/r-1$	$\frac{\frac{SSR}{r-1}}{\frac{SSE}{r(m-1)}}$	$P(F > F_0)$
SSE	$\sum_{i=1}^{r} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_i)^2$	$r(m-1)$	$SSE/r(m-1)$		

To compute efficiently, we always compute

$$SST = \sum_{i=1}^{r}\sum_{j=1}^{m}(y_{ij}-\bar{y})^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{m} y_{ij}^2 - \frac{T^2}{n}$$

$$T = \sum_{i,j} y_{ij} \qquad n = m * r$$

$$T^2 = \sum_{ij} y_{ij}^2$$

$$T_i = \sum_{j} y_{ij}$$

$$SSR = \sum_{i=1}^{r} m(\bar{y}_i - \bar{y})^2$$

$$= \sum_{i=1}^{r} m\left(\frac{T_i}{m} - \frac{T}{n}\right)^2$$

$$= \sum_{i=1}^{r} \frac{T_i^2}{m} - 2\cdot\frac{T_i \cdot T}{n} + m\cdot\frac{T^2}{n^2}$$

$$= \frac{1}{m}\sum_{i=1}^{r} T_i^2 - \frac{T^2}{n}$$

$\boxed{\text{ANOVA}}$ → actually one special realization of F-test in conventional LR model

$\boxed{\text{Terminology:}}$ ① $\boxed{x_1, \dots, x_p}$ is Continuous variables

test $p \in S$, $\beta_p = 0$ ⇒ $\boxed{\text{F-test}}$

$\boxed{\text{Constraint LR Model}}$

$$\frac{\frac{SSE(H_0) - SSE}{df_1}}{\frac{SSE}{n-p-1}} \sim F$$

☆ $$\frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F$$ is one specific form

② $\text{Null} \subseteq A \subseteq A + A{:}B$ .... ⇝ $\boxed{\text{Nested Factors}}$

A, B ⇝ Factors (discrete variables)

$$\boxed{\begin{aligned} \text{test: } & \beta_{A2} = \beta_{A2} = \cdots = \beta_{Ap_A} = 0 \\ & (\beta_{B11} = \cdots = \beta_{B1p_B} = \cdots = \beta_{Bp_Ap_B} = 0) \end{aligned}}$$

→ principly, apply F-test in LR Framework !

→ $\boxed{\text{ANOVA}}$

Remark : 1) ① and ② are actually equivalent , while ② have the "nested" constraints !

2) From ② , we care more about the quanlitative conclusion . That is, whether these factors really contribute to our predicted target (hypothesis testig resnlt)

→ in ANOVA, this can be done $\boxed{\text{without}}$ actually going through parameters estimation

$$\boxed{\begin{aligned} & RSS \leftrightarrow SSE \\ & (y_{ij} - \hat{y}_{ij})^2 \end{aligned}}$$

[E.g.] $\underbrace{\text{Null}}_{M_1} \subseteq \underbrace{A}_{M_2} \subseteq \underbrace{A + A{:}B}_{M_3}$

one-way [M_1 vs M_2]

$$\boxed{\begin{aligned} & RSS(\text{Null}) - RSS(M_1) \\ & = (\bar{y}_i - \bar{y})^2 \end{aligned}}$$

$$\boxed{\begin{aligned} & \text{Test } A{:}B \text{ significance} \\ & SSR = RSS(M_2) - RSS(M_3) \\ & SSE = RSS(M_3) \end{aligned}}$$