Lecture 17

§1 ANOVA中的 statistics

1. Definition: total sum of square (TSS / SST) (总平方和统计量)

TSS 可被表示为

$$TSS = \mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}^T J \mathbf{y}$$
$$= \mathbf{y}^T(I-\frac{1}{n}J)\mathbf{y}$$

其中 J 为所有元素均为 1 的矩阵

证明:

$$TSS = \sum_{i=1}^{n}(y_i-\bar{y})^2$$
$$= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$= \mathbf{y}^T\mathbf{y} - (\frac{1}{n}J\mathbf{y})^T(\frac{1}{n}J\mathbf{y}) \quad (\frac{1}{n}J\mathbf{y} = [\bar{y},\bar{y},\cdots,\bar{y}]^T = \bar{y}\cdot\mathbf{1})$$
$$= \mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}J\mathbf{y} \quad (\frac{1}{n}J \text{ 为 idempotent}: \frac{1}{n}J\cdot\frac{1}{n}J = \frac{1}{n}J)$$

2. Property: $I-\frac{1}{n}J$ 的性质

① $I-\frac{1}{n}J$ 为 symmetric matrix, 即 $(I-\frac{1}{n}J)^T = I-\frac{1}{n}J$

证明:

$$(I-\frac{1}{n}J)^T = I^T - \frac{1}{n}J^T = I-\frac{1}{n}J$$

注: 若 matrix A 为 symmetric, 则 $\mathbf{y}^T A\mathbf{y}$ 为 quadratic form, 因此 TSS 为 quadratic form

② $I-\frac{1}{n}J$ 为 idempotent matrix, 即 $(I-\frac{1}{n}J)(I-\frac{1}{n}J) = I-\frac{1}{n}J$

证明:

$$(I-\frac{1}{n}J)(I-\frac{1}{n}J) = I-\frac{2}{n}J+\frac{1}{n}J\frac{1}{n}J = I-\frac{2}{n}J+\frac{1}{n}J = I-\frac{1}{n}J$$

③ $I-\frac{1}{n}J$ 的 rank 为 $n-1$ (为 SST 的 d.f.)

证明:

$$\text{rank}(I-\frac{1}{n}J) = \text{rank}(I) - \text{rank}(\frac{1}{n}J) = n-1$$

3. Definition: sum of square errors (SSE / RSS) (误差/残差平方和)

SSE 可被表示为

$$SSE = \hat{\mathbf{e}}^T\hat{\mathbf{e}}$$
$$= \mathbf{y}^T(I-H)\mathbf{y}$$

其中 J 为所有元素均为 1 的矩阵

证明:

$$SSE = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$
$$= \hat{\mathbf{e}}^T\hat{\mathbf{e}}$$
$$= [(I-H)\mathbf{y}]^T[(I-H)\mathbf{y}]$$
$$= \mathbf{y}^T(I-H)^T(I-H)\mathbf{y}$$
$$= \mathbf{y}^T(I-H)\mathbf{y} \quad (I-H \text{ 为 symmetric 且 idempotent})$$

注：SSE 为 quadratic form，$rank(I-H) = rank(I) - rank(H) = n - \sum_{i=1}^{n} h_{ii} = n-2$.

因此 $d.f.(SSE) = n-2$

4. Definition: sum of squared due to the regression model (SSR / SSReg)（回归平方和）

SSR 可被表示为

$$SSR = \mathbf{y}^T H \mathbf{y} - \frac{1}{n}\mathbf{y}^T J \mathbf{y}$$
$$= \mathbf{y}^T (H - \frac{1}{n}J)\mathbf{y}$$

证明：

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$$
$$= \sum_{i=1}^{n} (\hat{y}_i^2 - 2\hat{y}_i\bar{y}_i + \bar{y}_i^2)$$
$$= \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y} \quad \left(由于 \sum_{i=1}^{n}\hat{y}_i = \sum_{i=1}^{n} y_i，有 \sum_{i=1}^{n}(-2\hat{y}_i\bar{y}_i + \bar{y}_i^2) = -2n\bar{y}_i^2 + n\bar{y}_i^2 = -n\bar{y}_i^2\right)$$
$$= \hat{\mathbf{y}}^T\hat{\mathbf{y}} - \frac{1}{n}\mathbf{y}^T J \mathbf{y}$$
$$= (H\mathbf{y})^T(H\mathbf{y}) - \frac{1}{n}\mathbf{y}^T J \mathbf{y}$$
$$= \mathbf{y}^T H \mathbf{y} - \frac{1}{n}\mathbf{y}^T J \mathbf{y} \quad (H 为 symmetric 且 idempotent)$$

注：SSR 为 quadratic form，$rank(H - \frac{1}{n}J) = rank(H) - rank(\frac{1}{n}J) = 2 - 1 = 1$

因此 $d.f.(SSR) = 1$

5. Lemma: TSS = SSE + SSR（Cochran's theorem）

若 observations $(X, \mathbf{y})$ 服从 linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

其中 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$

定义 $\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}}$ (fitted value)，则有

$$TSS = SSE + SSR$$

即 $\mathbf{y}^T(I - \frac{1}{n}J)\mathbf{y} = \mathbf{y}^T(I-H)\mathbf{y} + \mathbf{y}^T(H - \frac{1}{n}J)\mathbf{y}$

证明：（第一种证法）

$$TSS = \mathbf{y}^T(I - \frac{1}{n}J)\mathbf{y}$$
$$= \mathbf{y}^T(I - H + H - \frac{1}{n}J)\mathbf{y}$$
$$= \mathbf{y}^T(I-H)\mathbf{y} + \mathbf{y}^T(H - \frac{1}{n}J)\mathbf{y}$$
$$= SSE + SSR$$

证明：（第二种证法）

由于 $TSS = \mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}^T J \mathbf{y}$，先处理 $\mathbf{y}^T\mathbf{y}$：

$$\mathbf{y}^T\mathbf{y} = (\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}})$$
$$= (\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (\mathbf{y} - X\hat{\boldsymbol{\beta}})^T X\hat{\boldsymbol{\beta}} + (X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}) + (X\hat{\boldsymbol{\beta}})^T X\hat{\boldsymbol{\beta}}$$
$$= \hat{\mathbf{e}}^T\hat{\mathbf{e}} + \hat{\mathbf{e}}^T X\hat{\boldsymbol{\beta}} + (X\hat{\boldsymbol{\beta}})^T\hat{\mathbf{e}} + \hat{\boldsymbol{\beta}} X^T X\hat{\boldsymbol{\beta}}$$

其中，

① $\hat{\mathbf{e}}^T X\hat{\boldsymbol{\beta}} = (X\hat{\boldsymbol{\beta}})^T\hat{\mathbf{e}} = 0$ 　由于 $\hat{\mathbf{e}}^T X = (\mathbf{y} - H\mathbf{y})^T X = \mathbf{y}^T X - \mathbf{y}^T X = \mathbf{0}$

② $\hat{\beta}X^TX\hat{\beta} = y^THy$ 由于 $X\hat{\beta} = \hat{y} = Hy$

因此，

$$y^Ty = \hat{e}^T\hat{e} + y^THy$$

$$TSS = \underbrace{\hat{e}^T\hat{e}}_{SSE} + \underbrace{y^THy - \frac{1}{n}y^TJy}_{SSR}$$

$$= SSE + SSR$$

6. Property: SSE/RSS 的期望（用SSE估计 σ^2）

对于 $SSE = \hat{e}^T\hat{e} = y^T(I-H)y$，有

$$E(SSE) = (n-2)\sigma^2$$

证明:

$$\begin{aligned}
E(SSE) &= E(\hat{e}^T\hat{e}) \\
&= E(y^T(I-H)y) \\
&= E(\mathrm{trace}(y^T(I-H)y)) \quad \text{(由于 } y^T(I-H)y \text{ 为 scalar)} \\
&= E(\mathrm{trace}((I-H)yy^T)) \quad \text{(由于 } \mathrm{trace}(AB) = \mathrm{trace}(BA)\text{)} \\
&= \mathrm{trace}[(I-H)E(yy^T)] \\
&= \mathrm{trace}[(I-H)(\sigma^2 I + X\beta\beta^TX^T)] \quad (E(yy^T) = Var(y) + E(y)E(y^T)) \\
&= \mathrm{trace}[(I-H)\sigma^2 + X\beta\beta^TX^T - X(X^TX)^{-1}X^TX\beta\beta^TX^T] \\
&= \mathrm{trace}(I-H)\sigma^2 \\
&= (n-2)\sigma^2
\end{aligned}$$

注：由此证明了 $S^2 = \frac{RSS}{n-2}$ 为 σ^2 的无偏估计量

7. R中的 ANOVA

The anova command is one way in R to produce an ANOVA table, in addition to analysing it. For example, for the 654-point SLR problem in Assignment 2, question 1:

```
a2 = read.table("data.txt",sep=" ",header=T) # Load the data set
fev <- a2$fev; age <- a2$age
mod1 = lm(fev~age)
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: fev
##            Df  Sum Sq  Mean Sq  F value    Pr(>F)
## age         1  280.92  280.919   872.18  < 2.2e-16 ***
## Residuals 652  210.00    0.322
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```