One-Way ANOVA Model

group 1	$y_{11},$	y_{12}	y_{13} $\cdots$	y_{1n_1}
group 2	$y_{21},$	y_{22}	$\cdots$	y_{2n_2}
	$\cdots\cdots$			
group g	$y_{g1},$	$y_{g2},$	$\cdots$	y_{gn_g}

g is # of groups,

n_i denotes # of obs in the i-th group,

and the total sample size $n = \sum_{i=1}^{g} n_i$.

Model Equation

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad e_{ij} \text{ iid } \sim \mathsf{N}(0, \sigma^2).$$

The unknown parameters are

$$(\mu, \quad \alpha_1, \quad \ldots, \quad \alpha_g).$$

Matrix Form

For simplicity, consider a simple case $g = 3, n_1 = 3$, $n_2 = 2$ and $n_3 = 2$. Write the one-way ANOVA model in matrix form

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \text{err.}$$

The model is over-parameterized, i.e, columns of $\mathbf{X}$ are linearly dependent.

Projection Matrix

Let $\mathbf{Z}$ denote the design matrix $\mathbf{X}$ without the 1st column. Is it easy to show that $C(\mathbf{Z}) = C(\mathbf{X})$ and $\mathbf{Z}$ is of full rank. So the projection matrix can be computed via $\mathbf{Z}$:

$$\mathbf{H} = \mathbf{Z}(\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^t.$$

For the $g = 3$ example, we have

$$\mathbf{H}_{n\times n} = \begin{pmatrix} \frac{1}{3}\mathbf{J}_3^3 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{J}_2^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{2}\mathbf{J}_2^2 \end{pmatrix},$$

where $\mathbf{J}_m^m$ denotes an $m \times m$ matrix with all entries being 1.

- The LS fit for y_{ij} is the corresponding group mean

$$\hat{y}_{ij} = \bar{y}_{i\cdot}$$

- Residuals

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

- RSS

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{i\cdot} \right)^2,$$

  i.e., the within-group variation.

Comparing Nested Modesl and F-tests

- Are all groups really different? State the hypothesis in terms of models

$$H_a \quad : \quad y_{ij} = \mu + \alpha_i + e_{ij}$$

$$H_0 \quad : \quad y_{ij} = \mu + e_{ij}$$

- They are two nested models, then we can use F-test.

$$\frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(g-1)}{\mathsf{RSS}_a/(n-g)} \sim F_{g-1,n-g},$$

  under H_0. The test statistic can also written as

$$\frac{\sum_{i=1}^{g} n_i(y_{i\cdot} - y_{\cdot\cdot})^2/(g-1)}{\sum_{i,j}(y_{ij} - y_{i\cdot})^2/(n-g)} = \frac{\text{Between-group Variation}/(g-1)}{\text{Within-group Variation}/(n-g)}.$$

- How to use F-test to decide whether we should merge some groups?

LS Coefficient Estimation

As mentioned before, the LS estimates for $(\mu, \alpha_1, \ldots, \alpha_g)$ are not unique since the design matrix is not of full rank.

If you fit a one-way ANOVA model using `lm`, `R` will return you a g-dim regression coefficient vector. The LS estimates returned by `R` vary depending on the `contrast` option.

Since the one-way ANOVA model is over-parameterized, to obtain an estimate of the coefficients, we need to put some constraint on μ or α_i's.

- $\mu = 0$: What's the design matrix $\mathbf{X}$? How to interpret the parameters? (the default case; `lm` $\sim$ `X -1`)
- $\alpha_1 = 0$: What's the design matrix $\mathbf{X}$? How to interpret the parameters? (`contr.treatment`)
- $\sum \alpha_i = 0$: What's the design matrix $\mathbf{X}$? How to interpret the parameters? (`contr.sum`)
- Suffices to remember the default case; the interpretations are not important.

$\mu = 0$ (i.e., regression without the intercept):

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \text{err} = \begin{pmatrix} \alpha_1 \mathbf{1}_3 \\ \alpha_2 \mathbf{1}_2 \\ \alpha_3 \mathbf{1}_2 \end{pmatrix} + \text{err}.$$

So α_i denotes the (population) mean of group i.

`contr.treatment` ($\alpha_1 = 0$)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \text{err} = \begin{pmatrix} \mu \mathbf{1}_3 \\ (\mu + \alpha_2)\mathbf{1}_2 \\ (\mu + \alpha_3)\mathbf{1}_2 \end{pmatrix} + \text{err}.$$

So μ denotes the (population) mean of group 1 and α_2 denotes the difference between the mean of group 1 and mean of group 2, etc.

`contr.sum` ($\sum_i \alpha_i = 0$)

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \text{err} = \begin{pmatrix} (\mu + \alpha_1)\mathbf{1}_3 \\ (\mu + \alpha_2)\mathbf{1}_2 \\ (\mu - \alpha_1 - \alpha_2)\mathbf{1}_2 \end{pmatrix} + \text{err}.$$

So μ denotes the average of the group (population) means, and α_i denotes the difference between the mean of group i and μ.

Contrasts

- A linear combination of the group means $\sum_{i=1}^{g} c_i \alpha_i$ is called a contrast if $\sum_i c_i = 0$.
  - $\alpha_1 - \alpha_2$: $c_1 = 1$, $c_2 = -1$, and other c_i's $= 0$.
  - $(\alpha_1 + \alpha_2)/2 - \alpha_3$: $c_1 = c_2 = 1/2$, $c_3 = -1$, and other c_i's $= 0$.
- The LS estimate (i.e., BLUE) of $\sum_{i=1}^{g} c_i \alpha_i$ is $\sum_{i=1}^{g} c_i \bar{y}_{i\cdot}$ with s.e. $\hat{\sigma}\sqrt{\sum_i c_i^2/n_i}$.
- The (individual) $(1-\alpha)$ CI for $\sum_{i=1}^{g} c_i \alpha_i$ is given by

$$\sum_i c_i \bar{y}_{i\cdot} \pm t_{n-g}^{\alpha/2} \hat{\sigma} \sqrt{\sum_i \frac{c_i^2}{n_i}}.$$

Decompose BSS Using Orthogonal Contrasts

There is a one-to-one correspondence between the following two subspaces: each has dimension $(g-1)$, but one is from $\mathbb{R}^g$ and one is from $\mathbb{R}^n$.

- The contrast space $\{\mathbf{c} \in \mathbb{R}^g : c_1 + \cdots + c_g = 0.\}$.
- $\mathcal{M}_{A.0} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \in \mathcal{M}_A \text{ and } \mathbf{x} \perp \mathcal{M}_0\}$ where $\mathcal{M}_A$ denotes the estimation space for one-way ANOVA model and $\mathcal{M}_0$ denotes the estimation space for an intercept-only model.

Vectors in $\mathcal{M}_{A.0}$ take the following form

$$\begin{pmatrix} b_1\mathbf{1}_{n_1} \\ b_2\mathbf{1}_{n_2} \\ \cdots \\ b_g\mathbf{1}_{n_g} \end{pmatrix}, \qquad \text{where } \sum_{i=1}^{g} n_i b_i = 0.$$

Apparently, this vector above corresponds to a contrast vector $(n_1 b_1, \ldots, n_g b_g)^t$.

For any contrast vector $\mathbf{c}$, the corresponding LS estimate of $\mathbf{c}^t\boldsymbol{\alpha}$ is given by

$$\mathbf{a}^t\mathbf{y} = \left(\frac{c_1}{n_1}\mathbf{1}_{n_1}^t, \ldots, \frac{c_g}{n_g}\mathbf{1}_{n_g}^t\right)\mathbf{y},$$

and vector $\mathbf{a} \in \mathcal{M}_{A.0}$.

Let $\mathbf{a}_1, \ldots, \mathbf{a}_{g-1}$ be an orthogonal basis for $\mathcal{M}_{A.0}$, then the norm-square of any vector projected to $\mathcal{M}_{A.0}$ can be decomposed as sum of norm-square of its projection to each one-dim space spanned by $\mathbf{a}_j$:

$$\mathsf{SS}(\mathcal{M}_{A.0}) = \mathsf{SS}_1 + \cdots + \mathsf{SS}_{g-1}, \quad \mathsf{SS}_j = (\mathbf{y}^t \mathbf{a}_j)^2 / \|\mathbf{a}_j\|^2.$$

Define orthogonal contrasts to be two contrast vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ such that $\sum_i c_{i1} c_{i2} / n_i = 0$. For balanced design, it means $\mathbf{c}_1$ and $\mathbf{c}_2$ are also orthogonal as two g-dim vectors.

Let $\mathbf{c}_1, \ldots, \mathbf{c}_{g-1}$ be a set of (pairwise) orthogonal contrasts. For each $\mathbf{c}_i$, define $\mathbf{a}_j = (\frac{c_{j1}}{n_1} \mathbf{1}_{n_1}^t, \ldots, \frac{c_{jg}}{n_g} \mathbf{1}_{n_g}^t)^t$. Then $(\mathbf{a}_1, \ldots, \mathbf{a}_{g-1})$ form an orthogonal basis for $\mathcal{M}_{A.0}$, and

$$\mathsf{SS}_j = (\mathbf{y}^t \mathbf{a}_j)^2 / \|\mathbf{a}_j\|^2 = \frac{\left(\sum_i \bar{y}_{i\cdot} c_{ji} \right)^2}{\sum_i c_{ji}^2 / n_i}.$$

Which contrast has the largest SS?