16 One-Way Analysis of Variance

16.1 The Fixed Effects Model

The one-way ANOVA model is a linear model given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, i. e.

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_n \end{pmatrix} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0}_{n_1} & \cdots & \mathbf{0}_{n_1} \\ \mathbf{0}_{n_2} & \mathbf{1}_{n_2} & \cdots & \mathbf{0}_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_p} & \mathbf{0}_{n_p} & \cdots & \mathbf{1}_{n_p} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix},$$

with $n = n_1 + \ldots n_p$. We assume $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Note that $\operatorname{rank}(\mathbf{X}_{n\times p}) = p$ is of full rank. We want to test the hypothesis $H : \mu_1 = \mu_2 = \cdots = \mu_p$. This can be written as a linear hypothesis $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, i. e.

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Note that $\operatorname{rank}(\mathbf{A}) = p - 1$. From Theorem 11.8 we have that the test statistic for testing $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ is

$$F = \frac{(RSS_H - RSS)/(p-1)}{RSS/(n-p)}$$

with $F \sim F_{p-1,\, n-p}$ if H is true.

16.1 Theorem: The least squares estimates for μ_i is $\bar{Y}_i = \sum_j Y_{ij}/n_i$, and therefore

$$RSS = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2.$$

16.2 Note: The term $RSS = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$ is called the within group sums of squares. From Theorem 6.13 we have that $RSS/\sigma^2 \sim \chi^2_{n-p,}$ and hence $E[RSS/(n-p)] = \sigma^2$. The term $RSS/(n-p)$, an unbiased estimate of σ^2, is called within group mean squares.

16.3 Theorem: Under H we have $\mu_1 = \ldots = \mu_p = \mu$, and the least squares estimate for μ is $\bar{Y} = \sum_i \sum_j Y_{ij}/n$. We therefore have

$$RSS_H = (\mathbf{Y} - \hat{\mathbf{Y}}_H)'(\mathbf{Y} - \hat{\mathbf{Y}}_H) = \sum_i \sum_j (Y_{ij} - \bar{Y})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2 + \sum_i \sum_j (\bar{Y}_i - \bar{Y})^2.$$

From the above it follows that

$$RSS_H - RSS = \sum_i \sum_j (\bar{Y}_i - \bar{Y})^2 = \sum_i n_i (\bar{Y}_i - \bar{Y})^2.$$

16.4 Note: The term $RSS_H - RSS = \sum_i n_i(\bar{Y}_i - \bar{Y})^2$ is called between group sums of squares. From Theorem 11.6 it follows that if H is true, $(RSS_H - RSS)/\sigma^2 \sim \chi^2_{p-1}$, and $E[(RSS_H - RSS)/(p-1)] = \sigma^2$. The term $(RSS_H - RSS)/(p-1)$ is therefore another unbiased estimate of σ^2, called between group mean squares. Note that in Theorem 11.8 we showed that the between group mean squares are independent of the within group mean squares!

16.5 Note: The above is usually summarized in an ANOVA table:

source	sum of squares	df	mean square	test statistic
between treatments	$\text{SS}_T = \sum_i n_i(\bar{Y}_i - \bar{Y})^2$	$p-1$	$\text{MS}_T = \text{SS}_T/(p-1)$	$F = \text{MS}_T/\text{MS}_R$
within treatments	$\text{SS}_R = \sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2$	$n-p$	$\text{MS}_R = \text{SS}_R/(n-p)$	
total	$\sum_i \sum_j (Y_{ij} - \bar{Y})^2$	$n-1$		

16.6 Note: If H is false, we know from Theorem 11.15 that $F \sim F_{p-1,n-p}(\lambda)$, with

$$\sigma^2 2\lambda = \boldsymbol{\mu}'(\mathbf{P}_\Omega - \mathbf{P}_\omega)\boldsymbol{\mu} = (RSS_H - RSS)\,|_{\mathbf{Y}=\boldsymbol{\mu}} = \sum_i \sum_j (\mu_i - \mu)^2 = \sum_i n_i(\mu_i - \mu)^2.$$

From Theorem 11.6 follows that

$$E[\text{MS}_T] = E[(RSS_H - RSS)/(p-1)] = \sigma^2 + \frac{\sum_i n_i(\mu_i - \mu)^2}{p-1}.$$

Note that $E[\text{MS}_R] = E[RSS/(n-p)] = \sigma^2$, whether or not H is true.

16.2 The Random Effects Model

So far, we viewed models as conditional on a structure S. For example, by assuming fixed μ_i in the above, we have used a model defined by the structure

$$E(Y_{ij}|S) = \mu_i \quad \text{and} \quad \text{cov}\,(Y_{ij}, Y_{i'j'}|S) = \begin{cases} \sigma^2 & (i',j') = (i,j) \\ 0 & \text{elsewhere.} \end{cases}$$

If however we entertain a model which specifies that the μ_i are a sample from a population of possible groups which could have been selected for inclusion in the experiment, we have an entirely different situation. We must specify the stochastic structure of the μ_i and then take expectations with respect to this structure to obtain the expected values of the sums of squares.

One such model is the so called random effects model in which we assume that

$$E(\mu_i) = \mu \;\text{ and }\; \text{cov}\,(\mu_i, \mu_{i'}) = \begin{cases} \sigma_1^2 & i' = i \\ 0 & \text{elsewhere.} \end{cases}$$

16.7 Theorem: This model implies that the Y_{ij} are correlated since

$$\text{cov}\,(Y_{ij}, Y_{i'j'}) = \begin{cases} \sigma^2 + \sigma_1^2 & (i', j') = (i, j) \\ \sigma_1^2 & i' = i, j' \neq j \\ 0 & \text{elsewhere.} \end{cases}$$

The correlation between any two observations in the same group is given by

$$\text{corr}\,(Y_{ij}, Y_{ij'}) = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2},$$

which is called the intra-class correlation.

16.8 Theorem: For the random effects model we have:

(a) $E[\text{MS}_T] = \sigma^2 + \frac{1}{p-1}\left(n - \sum_i \frac{n_i^2}{n}\right)\sigma_1^2$,

(b) $E[\text{MS}_R] = \sigma^2$.

16.9 Note: If we have a random effects model with $n_1 = \ldots = n_p = r$, then $E[\text{MS}_T] = \sigma^2 + r\sigma_1^2$.

16.10 Theorem: If in the random effects model we have $n_1 = \ldots = n_p = r$, then:

(a) $\text{SS}_T/(\sigma^2 + r\sigma_1^2) \sim \chi^2_{p-1}$,

(b) $\text{SS}_R/\sigma^2 \sim \chi^2_{p(r-1)}$,

(c) SS_T and SS_R are independent.

16.11 Note: We can test $H : \sigma_1^2 = 0$ versus $\sigma_1^2 > 0$ using the test statistic $F = \text{MS}_T/\text{MS}_R$, with $F \sim F_{p-1,\, p(r-1)}$ in the balanced case if H is true. If we reject H, we estimate the variance components as:

$$\widehat{\sigma^2} = \text{MS}_R \qquad \text{and} \qquad \widehat{\sigma_1^2} = (\text{MS}_T - \text{MS}_R)/r.$$

16.12 Note: Since $E[Y_{ij}] = \mu$, it follows that $\bar{Y}$ is an unbiased estimator of μ. However, the presence of correlation between the Y_{ij} in the random effects model does not imply that $\bar{Y}$ has any optimality properties.

16.13 Theorem: The best linear unbiased estimate (BLUE) for μ is

$$\widehat{\mu} = \frac{1}{\sum_{j=1}^{p}\left(\frac{n_j}{\sigma^2+n_j\sigma_1^2}\right)} \sum_{i=1}^{p} \frac{n_i}{\sigma^2+n_i\sigma_1^2}\bar{Y}_i.$$

16.14 Note: In the balanced case with $n_1 = \ldots = n_p = r$, the above reduces to $\widehat{\mu} = \bar{Y}$.

16.15 Theorem: For the fixed effects model we have

$$L(\mu_1,\ldots,\mu_p,\sigma^2|\mathbf{Y}) = (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{p} n_i(\bar{Y}_i-\mu_i)^2\right\}.$$

16.16 Note: The sufficient statistics for $\mu_1,\ldots,\mu_p$ and σ^2 are $\bar{Y}_1,\ldots,\bar{Y}_p$ and $\sum_i\sum_j(Y_{ij}-\bar{Y}_i)^2$.

16.17 Theorem: For the random effects model we have

$$\begin{aligned} L(\mu,\sigma^2,\sigma_1^2|\mathbf{Y}) &= (2\pi)^{-\frac{n}{2}}(\sigma^2)^{-\frac{n-p}{2}}\left(\prod_{i=1}^{p}\sigma^2+n_i\sigma_1^2\right)^{-\frac{1}{2}} \times \\ &\quad \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2 - \frac{1}{2}\sum_{i=1}^{p}\frac{n_i(\bar{Y}_i-\mu)^2}{\sigma^2+n_i\sigma_1^2}\right\}. \end{aligned}$$

16.18 Note: The sufficient statistics for μ,σ^2 and σ_1^2 are $\bar{Y}_1,\ldots,\bar{Y}_p$ and $\sum_i\sum_j(Y_{ij}-\bar{Y}_i)^2$ as well.

16.19 Note: In the balanced case with $n_1 = \ldots = n_p = r$, we have

$$\sum_{i=1}^{p}\frac{n_i(\bar{Y}_i-\mu)^2}{\sigma^2+n_i\sigma_1^2} = \frac{r\sum_{i=1}^{p}(\bar{Y}_i-\mu)^2}{\sigma^2+r\sigma_1^2} = \frac{r\sum_{i=1}^{p}(\bar{Y}_i-\bar{Y})^2+rp(\bar{Y}-\mu)^2}{\sigma^2+r\sigma_1^2},$$

and the sufficient statistics for μ,σ^2 and σ_1^2 are $\bar{Y}$, $\sum_i\sum_j(Y_{ij}-\bar{Y}_i)^2$, and $\sum_i(\bar{Y}_i-\bar{Y})^2$.