Ch12. Analysis of Variance: Unbalanced Data

1 One-Way Model

The non-full-rank and cell means versions of the one-way unbalanced model are

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} := \mu_i + \epsilon_{ij}, \tag{1}$$

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n_i$. For making inferences, we assume ϵ_{ij}'s are independently distributed as $N(0, \sigma^2)$.

1.1 Estimation and Testing

To estimate the μ_i's, we begin by writing the $N = \sum_i n_i$ observations for the above model by

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\mu} + \boldsymbol{\epsilon},$$

for which the normal equation is given by

$$\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\mu}} = \boldsymbol{W}^T\boldsymbol{y}.$$

The resulting parameter estimator is given by

$$\hat{\boldsymbol{\mu}} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{y} = \bar{\boldsymbol{y}} = (\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, \ldots, \bar{y}_{k\cdot})^T,$$

where $\bar{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij}/n_i$.

To test $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, we compare the full model in (1) with the reduced model $y_{ij} = \mu + \epsilon^*_{ij}$. The difference $SS(\mu_1, \mu_2, \ldots, \mu_k) - SS(\mu)$ is denoted by

$$SSB = \hat{\boldsymbol{\mu}}'\boldsymbol{W}'\boldsymbol{y} - N\bar{y}^2_{\cdot\cdot} = \sum_{i=1}^{k} \frac{y^2_{i\cdot}}{n_i} - \frac{y^2_{\cdot\cdot}}{N} = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2,$$

for "between" sum of squares, and it has $k-1$ degrees of freedom.

Then the F-statistic for testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is given by

$$F = \frac{SSB/(k-1)}{SSE/(N-k)}.$$

If H_0 is true, F is distributed as $F(k-1, N-k)$.

1.2 Contrasts

A contrast is defined as $\delta = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$, where $\sum_{i=1}^k c_i = 0$. The contrast can be expressed as $\delta = \boldsymbol{c}'\boldsymbol{\mu}$. Then the F-statistic for testing $H_0 : \delta = 0$ is given by

$$F = \frac{(\boldsymbol{c}'\hat{\boldsymbol{\mu}})'[\boldsymbol{c}'(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{c}]^{-1}\boldsymbol{c}'\hat{\boldsymbol{\mu}}}{s^2} = \frac{(\sum_{i=1}^k c_i\bar{y}_{i\cdot})^2/(\sum_{i=1}^k c_i^2/n_i)}{s^2},$$

where $s^2 = SSE/(N-k)$.

If $H_0 : \delta = 0$ is true, then F is distributed as $F(1, N-k)$.

2 Two-Way Model

The unbalanced two-way model is given by

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \\ &= \mu_{ij} + \epsilon_{ijk}, \end{aligned} \tag{2}$$

where $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, b$, and $k = 1, 2, \ldots, n_{ij}$. The ϵ_{ijk}'s are assumed to be independently distributed as $N(0, \sigma^2)$. In this section, we consider the case in which all $n_{ij} > 0$.

2.1 Unconstrained Model

We first consider the unconstrained model in which the μ_{ij}'s are unrestricted. To illustrate the cell mean model (2), we use $a = 2$ and $b = 3$ with cell counts given by

$$n_{11} = 2, \quad n_{12} = 1, \quad n_{13} = 2, \quad n_{21} = 1, \quad n_{22} = 3, \quad n_{23} = 2,$$

for which $N = 11$. The cell model can be written as

$$\begin{aligned} y_{111} &= \mu_{11} + \epsilon_{111}, \\ y_{112} &= \mu_{11} + \epsilon_{112}, \\ &\vdots \\ y_{232} &= \mu_{23} + \epsilon_{232}. \end{aligned}$$

In matrix form, we have

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\mu} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \ldots, \mu_{23})^T$, and

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Since $\boldsymbol{W}$ is of full rank, we have

$$\hat{\boldsymbol{\mu}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y} = \bar{y},$$

by noting that $\boldsymbol{W}'\boldsymbol{W} = diag\{n_{11}, n_{12}, \ldots, n_{23}\}$.

For general a, b and N, an unbiased estimator of σ^2 is given by

$$s^2 = \frac{SSE}{\nu_E} = \frac{(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\mu}})'(\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\mu}})}{N - ab}.$$

Alternative forms of SSE are given by

$$SSE = \boldsymbol{y}'(I - \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}')\boldsymbol{y},$$

and

$$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij})^2.$$

Further, we have

$$s^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - 1)s_{ij}^2}{N - ab}.$$

We now construct tests for the general linear hypotheses $H_0 : \boldsymbol{a}'\boldsymbol{\mu} = 0$ and $H_0 : \boldsymbol{C}\boldsymbol{\mu} = 0$. The hypothesis $H_0 : \boldsymbol{a}'\boldsymbol{\mu} = 0$ can be tested using an F-statistic:

$$F = \frac{(\boldsymbol{a}'\hat{\boldsymbol{\mu}})'[\boldsymbol{a}'(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{a}]^{-1}(\boldsymbol{a}'\hat{\boldsymbol{\mu}})}{s^2} := \frac{SSA}{SSE/\nu_E},$$

which is distributed as $F(1, N - ab)$ if H_0 is true.

A test statistic for $H_0 : \boldsymbol{C}\boldsymbol{\mu} = 0$ is given by

$$F = frac(\boldsymbol{C}'\hat{\boldsymbol{\mu}})'[\boldsymbol{C}'(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{C}]^{-1}(\boldsymbol{C}'\hat{\boldsymbol{\mu}})/\nu_{AB}SSE/\nu_E,$$

which is distributed as $F(\nu_{AB}, \nu_E)$.

Note that different choice of $\boldsymbol{C}$ can represent the main effects and interaction effects of the two-way model. For example, the interaction hypothesis can be written as

$$H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23},$$

by noting

$$\mu_{11} - \mu_{12} = \mu + \alpha_1 + \beta_1 + \gamma_{11} - (\mu + \alpha_1 + \beta_1 + \gamma_{12}) = \beta_1 - \beta_2 + \gamma_{11} - \gamma_{12},$$

$$\mu_{21} - \mu_{22} = \beta_1 - \beta_2 + \gamma_{21} - \gamma_{22},$$

and

$$\mu_{31} - \mu_{32} = \beta_1 - \beta_2 + \gamma_{31} - \gamma_{32}.$$

Therefore equating them leads to testing

$$\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22} = 0, \quad \gamma_{21} - \gamma_{22} - \gamma_{31} + \gamma_{32} = 0.$$

The hypothesis $H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$ can be expressed as $H_0 : \boldsymbol{C}\mu = 0$, where

$$\boldsymbol{C} = \begin{pmatrix} 2 & -1 & -1 & -2 & 1 & 1 \\ 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix},$$

based on the formulation:

$$2(\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23}) = 0,$$

and

$$(\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23}) = 0.$$

2.2 Constrained Model

For constriants $\boldsymbol{G}\mu = 0$, the model can be expressed as

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\mu} + \boldsymbol{\epsilon} \quad \text{subject to } \boldsymbol{G}\boldsymbol{\mu} = 0.$$

One way to solve the model is to use the Langrange multiplier method. Alternatively, we can reparameterize the model using the matrix

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{K} \\ \boldsymbol{G} \end{pmatrix},$$

such that $\boldsymbol{A}$ is nonsingular under assumption that $\boldsymbol{G}$ is of full-row-rank.

Therefore,

$$\begin{aligned} \boldsymbol{y} = \boldsymbol{W}\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\epsilon} &= \boldsymbol{Z}\begin{pmatrix} \boldsymbol{K}\boldsymbol{\mu} \\ \boldsymbol{G}\boldsymbol{\mu} \end{pmatrix} + \boldsymbol{\epsilon}, \\ &:= (\boldsymbol{Z}_1, \boldsymbol{Z}_2)\begin{pmatrix} \boldsymbol{K}\boldsymbol{\mu} \\ \boldsymbol{G}\boldsymbol{\mu} \end{pmatrix} + \boldsymbol{\epsilon}, \\ &:= \boldsymbol{Z}_1\boldsymbol{\delta}_1 + \boldsymbol{Z}_2\boldsymbol{\delta}_2 + \boldsymbol{\epsilon}. \end{aligned}$$

Based on this reparameterization, the full-reduced-model approach can be applied to test $H_0 : \boldsymbol{\delta}_2 = \boldsymbol{G}\boldsymbol{\mu} = 0$.

3 Two-Way Model with Empty Cells

Consider the unbalanced two-way model in (2), but allow n_{ij} to be equal to 0 for one or more isolated cells. In some case, $\boldsymbol{W}$ is non-full-rank in that it has m columns equal to 0. In this case, the techniques of non-full-rank model we studied before can be applied for the analysis.

Consider an rearrangement of the matrix $\boldsymbol{W}$ such that

$$\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{0}),$$

that is, we assume that columns of $\boldsymbol{W}$ has been arranged with the columns of 0 occurring last. correspondingly, we have $\boldsymbol{\mu} = (\boldsymbol{\mu}_o^T, \boldsymbol{\mu}_e)^T$.

For such a model, we consider the reparameterization:

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{K} \\ \boldsymbol{G} \end{pmatrix}.$$

Then the following theorem holds:

Theorem 3.1 *Consider the constrained empty cells model with m empty cells. Partition $\boldsymbol{A}$ as*

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{K} \\ \boldsymbol{G} \end{pmatrix} = \begin{pmatrix} \boldsymbol{K}_1 & \boldsymbol{K}_2 \\ \boldsymbol{G}_1 & \boldsymbol{G}_2 \end{pmatrix}$$

conformal with the partitioned vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_o^T, \boldsymbol{\mu}_e^T)^T$. The elements of $\boldsymbol{\mu}$ are estimable (equivalently $\boldsymbol{Z}_1$ is full-rank) if and only if rank($\boldsymbol{G}_2$) $= m$.