## 20.1. Balanced One-Way Classification

*Cell means parametrization:*

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, I; j = 1, \ldots, J,$$

$$\varepsilon_{ij} \sim N(0, \sigma^2), \quad \text{independent.}$$

In matrix form, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, or

$$\begin{pmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_I \end{pmatrix} = \begin{pmatrix} \mathbf{1}_J & \mathbf{0}_J & \cdots & \mathbf{0}_J \\ \mathbf{0}_J & \mathbf{1}_J & \cdots & \mathbf{0}_J \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_J & \mathbf{0}_J & \cdots & \mathbf{1}_J \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_I \end{pmatrix},$$

where $\mathbf{Y}_i' = (Y_{i1}, \ldots, Y_{iJ})$, $\boldsymbol{\varepsilon}_i' = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$.

The least-squares estimates are:

$$\hat{\mu}_i = \bar{Y}_i.$$

So the $RSS$ is

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

## Alternative Parametrization of the Model:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

$\mu$ is interpreted as the overall mean ("grand mean"), $\alpha_i$ is interpreted as the difference between the mean of group $i$ and the overall mean.

Are the $\alpha_i$ estimable with this parametrization? No.

Note that $\alpha_i$ is like $\mu_i - \bar{\mu}$ from the previous parameterization. Therefore, $\sum_{i=1}^{I} \alpha_i = 0$ is a natural constraint because

$$\sum_{i=1}^{I} \alpha_i = \sum_{i=1}^{I} (\mu_i - \bar{\mu}) = 0$$

is a natural constraint. This is an *identifiability constraint* for the model.

The *unique* least-squares estimates satisfying $\sum \alpha_i = 0$ can be derived by the following trickery:

Rewrite $\varepsilon_{ij}$ as:

$$\varepsilon_{ij} = \bar{\varepsilon}_{\cdot\cdot} + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}) + (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot}).$$

where $\bar{\epsilon}_{i\cdot} = \frac{1}{J} \sum_j \epsilon_{ij}$ and $\bar{\epsilon}_{\cdot\cdot} = \frac{1}{IJ} \sum_i \sum_j \epsilon_{ij}$.

Square both sides of this expression and sum over all $i$ and $j$. The cross-product terms are 0.

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\varepsilon_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\bar{\varepsilon}_{\cdot\cdot}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}_{\cdot\cdot})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot})^2.$$

Now make some substitutions:

$$\begin{aligned}
\varepsilon_{ij} &= Y_{ij} - \mu - \alpha_i \\
\bar{\varepsilon}_{i\cdot} &= \bar{Y}_{i\cdot} - \mu - \alpha_i \\
\bar{\varepsilon}_{\cdot\cdot} &= \bar{Y}_{\cdot\cdot} - \mu - \frac{1}{I}\sum \alpha_i = \bar{Y}_{\cdot\cdot} - \mu
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \mu - \alpha_i)^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{\cdot\cdot} - \mu)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} - \alpha_i)^2 \\
&\quad + \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\cdot})^2.
\end{aligned}$$

What is the left hand side? RSS

This is minimized by the least-squares estimates:

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} \quad \text{and} \quad \hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}.$$

## 20.3. F-test for Group Differences

Test $H : \mu_1 = \mu_2 = \cdots = \mu_I$, i.e.,

$$H : \begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix} = \mathbf{0}.$$

The $F$ statistic is

$$F = \frac{(RSS_H - RSS)/(I-1)}{RSS/[IJ - I]},$$

where

$$RSS_H = \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

We have a similar algebraic identity:

$$\begin{aligned} \sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2 &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + \sum_i \sum_j (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ &= \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 + J \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\ RSS_H &= RSS + J \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \end{aligned}$$

The $F$ statistic becomes

$$F = \frac{J \sum_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2/(I-1)}{\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2/[IJ - I]} \sim F_{I-1, I(J-1)}.$$

The results of a one-way ANOVA are often displayed in an ANOVA table:

| Source | df | SS | MS | $F$ |
|---|---|---|---|---|
| Groups | $I-1$ | $SS = J\sum_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$ | $MS = \frac{SS_A}{I-1}$ | $\frac{MS_A}{MS_E}$ |
| Error | $I(J-1)$ | $SS_E = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$ | $MS_E = \frac{SS_E}{I(J-1)}$ | |
| Total | $IJ-1$ | $\sum_i \sum_j (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$ | | |

## 20.4. Example

Compare the breaking strength of steel plates for 5 brands of cars:

| Country | Brand | Mean |
|---|---|---|
| U.S. | GM(1) | $\mu_1$ |
| U.S. | GM(2) | $\mu_2$ |
| U.S. | Ford | $\mu_3$ |
| Japan | Toyota | $\mu_4$ |
| Japan | Honda | $\mu_5$ |

Data are measurements on $J = 4$ samples per brand. Suppose we get

| Source | df | SS | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| Groups | 4 | 10.24 | 2.56 | 2.56/0.34=7.53 | .0016 |
| Error | 15 | 5.10 | 0.34 | | |
| Total | 19 | 15.34 | | | |

Using the F-test, do we reject $H : \mu_1 = \mu_2 = \cdots = \mu_5$? Yes, we reject at the 0.05 and 0.01 level.

## 20.5. Orthogonal Contrasts

Some more algebra:

$$\begin{aligned}
\text{REG–SS} &= (\mathbf{PY})'(\mathbf{PY}) = \mathbf{Y}'\mathbf{PY} \\
&= \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \hat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} \\
&= \hat{\boldsymbol{\beta}}'\mathbf{A}'\mathbf{A}'^{-1}(\mathbf{X}'\mathbf{X})\mathbf{A}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}} \\
&= (\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}'^{-1}(\mathbf{X}'\mathbf{X})\mathbf{A}^{-1}](\mathbf{A}\hat{\boldsymbol{\beta}}) \\
&= (\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}})
\end{aligned}$$

This decomposition is general, but we now apply it to one-way ANOVA using a matrix $A$ whose rows are pairwise orthogonal (so that $\mathbf{A}'\mathbf{A}$ is diagonal). With the cell-means parameterization, $(\mathbf{X}'\mathbf{X})^{-1} = (1/J)\mathbf{I}$. Then

$$\begin{aligned}
\text{REG–SS} &= (\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}) \\
&= (\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}\mathbf{A}'/J]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}) \\
&= \sum(\mathbf{a}_i'\hat{\boldsymbol{\beta}})^2/[\mathbf{a}_i'\mathbf{a}_i/J]
\end{aligned}$$

In one-way ANOVA, we sometimes use *orthogonal contrasts* to further decompose the regression sums of squares.

## 20.6. Example Continued

Consider the following four orthogonal contrasts of the cell means. (There are four degrees of freedom, since the fifth degree of freedom is for the overall mean). We partition the "Groups" Sum-Of-Squares into four smaller Sums-of-Squares corresponding to these four orthogonal contrasts:

$$\begin{array}{ll} \text{U.S. vs. Japanese:} & \mathbf{a}_1'\boldsymbol{\beta} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{1}{2}, -\frac{1}{2})\boldsymbol{\beta} \\ \text{GM vs. Ford:} & \mathbf{a}_2'\boldsymbol{\beta} = (\frac{1}{2}, \frac{1}{2}, -1, 0, 0)\boldsymbol{\beta} \\ \text{GM(1) vs. GM(2):} & \mathbf{a}_3'\boldsymbol{\beta} = (1, -1, 0, 0, 0)\boldsymbol{\beta} \\ \text{Toyota vs. Honda:} & \mathbf{a}_4'\boldsymbol{\beta} = (0, 0, 0, 1, -1)\boldsymbol{\beta} \end{array}$$

Note that $\mathbf{a}_i'\mathbf{a}_j = 0, i \neq j$.

| Source | df | SS | MS | $F$ | $p$ |
|---|---|---|---|---|---|
| *Groups:* | 4 | 10.24 | 2.56 | 7.53 | .0016 |
| U.S. vs. Japanese | 1 | 4.39 | 4.39 | 12.91 | .0027 |
| GM vs. Ford | 1 | 1.69 | 1.69 | 4.97 | .041 |
| GM(1) vs. GM(2) | 1 | 2.64 | 2.64 | 7.76 | .014 |
| Toyota vs. Honda | 1 | 1.52 | 1.52 | 4.47 | .052 |
| Error | 15 | 5.10 | 0.331 | | |

## 20.7. Unbalanced Case

Suppose there are different numbers of observations per group:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, I; \; j = 1, \ldots, J_i.$$

Most of the same tricks from before still work. Rewrite $\varepsilon_{ij}$ as:

$$\varepsilon_{ij} = \bar{\varepsilon}_{\cdot\cdot} + (\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot}) + (\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot}).$$

Square both sides of this expression and sum over all $i$ and $j$. The cross-product terms are still 0.

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i}\varepsilon_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\bar{\varepsilon}_{\cdot\cdot}^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{\varepsilon}_{i\cdot} - \bar{\varepsilon}_{\cdot\cdot})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\varepsilon_{ij} - \bar{\varepsilon}_{i\cdot})^2.$$

Some substitutions are the same as before:

$$\begin{aligned} \varepsilon_{ij} &= Y_{ij} - \mu - \alpha_i \\ \bar{\varepsilon}_{i\cdot} &= \bar{Y}_{i\cdot} - \mu - \alpha_i \end{aligned}$$

But what about $\bar{\varepsilon}_{\cdot\cdot}$? It is NOT the average of the $\bar{\varepsilon}_{i\cdot}$.

$$\begin{aligned} \bar{\varepsilon}_{\cdot\cdot} &= \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\varepsilon_{ij} = \frac{1}{n}\sum_{i=1}^{I} J_i\bar{\varepsilon}_{i\cdot} = \frac{1}{n}\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \mu - \alpha_i) \\ &= -\mu + \frac{1}{n}\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \alpha_i) = \bar{Y}_{\cdot\cdot} - \mu - \frac{1}{n}\sum_{i=1}^{I} J_i\alpha_i \end{aligned}$$

To continue as we did before, the convenient identifiability constraint is now $\sum_{i=1}^{I} J_i\alpha_i = 0$. So then $\bar{\varepsilon}_{\cdot\cdot} = \bar{Y}_{\cdot\cdot} - \mu$ and the rest works out as before.

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\mu-\alpha_i)^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{Y}_{\cdot\cdot}-\mu)^2 \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot}-\alpha_i)^2 \\ &\quad + \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2.\end{aligned}$$

This shows that $\hat{\mu} = \bar{Y}_{\cdot\cdot}$ and $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$ when we use the identifiability constraint $\sum_{i=1}^{I} J_i\alpha_i = 0$.

Then the $F$ statistic for testing group differences is now

$$F = \frac{\sum_i J_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2/(I-1)}{\sum_i\sum_j(Y_{ij}-\bar{Y}_{i\cdot})^2/(n-I)} \sim F_{I-1,n-I}.$$

Thus the numerator of the $F$-statistic is still the "group sum of squares" (but now appropriately weighted).

Imbalance doesn't complicate things too much in one-way ANOVA. Not so for two-way ANOVA, etc.