Chapter 15: Introduction to the Design of Experimental and Observational Studies

The Models in Analysis of Variance (ANOVA) and in Regression are different. In regression model, all the response and predictors are continuous (quantitative) variables. However, in ANOVA model, the response is continuous but the predictors are categorical (qualitative) variables. There are some concepts here.

1. Factor and factor level. A factor is a predictor (explanatory or independent) variable. A factor level is a particular form of the factor. Mostly, the level can not be compared.

2. Single-factor and multi-factor studies. Single factor study means there is only one factor in the study. Thus, the model only includes one response and one predictor. Multi-factor means there are more than one factor in the study. An important case in the multi-factor study is two-way ANOVA model.

3. Experimental and Observational studies. An experimental study means the level of all the factors can be totally controlled. An observational study means the level of the factors can not be controlled. If some of them can be controlled and some of them can not, then people treat it as an experimental studies and called the controlled factor as a signal factor and the uncontrolled factor as a noise factor.

4. If a continuous variable is treated as a categorical variable, then it is also called a factor variable.

5. Treatment and block. The factor can be called either treatment or block variable. Both of them are treated as nominal variables. The interesting variable is called treatment and the uninteresting variables are called blocks.

6. Experimental design. If the factors levels can be controlled, then people try to make all the factors orthogonal to each other.

Chapter 16: One-Way (Single-factor) ANOVA Model

One Way ANOVA model has one response and one predictor. The response is continuous variable but the predictor is nominal variable.

**Mean Cells Model** Assume that factor A has I levels and in each level there are $n_i$ repeated observations. Then, the one-way ANOVA model can be expressed as

$$Y_{ij} = \mu_i + \epsilon_{ij},$$
i = 1, \cdots, I and j = 1, \cdots, n_i, n_i \geq 1, where Y_{ij} are response, \mu_i are parameters and 
\epsilon_{ij} \sim_{i.i.d.} N(0, \sigma^2) are error terms. Let

\[ n = \sum_{i=1}^{I} n_i. \]

**Note:** The book calls this model cell means model and the observation \( Y_{ij} \) the \( j \)-th trial
for the \( i \)-th factor level.

**Matrix Expression**

Let us denote \( 1_\ell \) to be the column vector of length \( \ell \) with all elements one, \( 0_\ell \) to be the
column vector of length \( \ell \) with all elements zero, and

\[
Y = \begin{pmatrix}
Y_{11} \\
Y_{12} \\
\vdots \\
Y_{1n_1} \\
\vdots \\
Y_{In_I}
\end{pmatrix},
\epsilon = \begin{pmatrix}
\epsilon_{11} \\
\epsilon_{12} \\
\vdots \\
\epsilon_{1n_1} \\
\vdots \\
\epsilon_{In_I}
\end{pmatrix},
\beta = \begin{pmatrix}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_I
\end{pmatrix},
\]

and

\[
X = \begin{pmatrix}
1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\
0_{n_2} & 1_{n_2} & \cdots & 0_{n_2} \\
\vdots & \vdots & \ddots & \vdots \\
0_{n_I} & 0_{n_I} & \cdots & 1_{n_I}
\end{pmatrix}.
\]

Then, the matrix expression for the cell means model is

\[ Y = X\beta + \epsilon. \]

It is clear that \( X \) is an \( n \times I \) matrix with rank \( I \).

**Comment:** People are interested in level means \( \mu_i \) sometimes, but mostly the mean
differences, \( \mu_i - \mu_{i_2} \), are an interesting issue.

**Parameter Estimates.** Use the least square method. Then, we have

\[ \hat{\beta} = (X'X)^{-1}X'Y. \]

It is clear that

\[
X'X = \begin{pmatrix}
n_1 & 0 & \cdots & 0 \\
0 & n_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & n_I
\end{pmatrix}
\]
\[(X'X)^{-1}X'Y = \begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ \vdots \\ \bar{Y}_I \end{pmatrix},\]

where 
\[
\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.
\]

Therefore, 
\[
\hat{Y}_{ij} = \hat{\mu}_i = \bar{Y}_i \sim N(\mu_i, \frac{1}{n_i}\sigma^2)
\]

and 
\[
\hat{\sigma}^2 = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \sim \frac{\sigma^2}{n-I} \chi^2_{n-I}.
\]

**Analysis of Variance.** Let us define sum of squares of total 

\[
SSTO = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}.)^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i + \bar{Y}_i - \bar{Y}.)^2
\]

\[
= \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{I} n_i (\bar{Y}_i - \bar{Y}.)^2.
\]

Therefore, we define 
\[
SSTR = \sum_{i=1}^{I} n_i (\bar{Y}_i - \bar{Y}.)^2,
\]

and 
\[
SSE = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.
\]

It is clear that 
\[
SSE \sim \sigma^2 \chi^2_{n-I}.
\]

Let us consider 
\[
E(SSTR) = \sum_{i=1}^{I} n_i E(\bar{Y}_i - \bar{Y}.)^2 = \sum_{i=1}^{I} n_i E[\bar{Y}_i^2 - 2\bar{Y}_i \bar{Y}. + \bar{Y}.^2]
\]

\[
= \sum_{i=1}^{I} n_i E[\frac{\sigma^2}{n_i} + \mu_i^2 + \frac{\sigma^2}{n} + \mu_.^2 - 2\mu_i \mu_. - 2\frac{\sigma^2}{n}]
\]

\[
=(I-1)\sigma^2 + \sum_{i=1}^{I} n_i (\mu_i^2 + \mu_.^2 - 2\mu_i \mu_.)
\]

\[
=(I-1)\sigma^2 + \sum_{i=1}^{I} n_i (\mu_i - \mu_.)^2.
\]
Table 1: One-Way ANOVA table

<table>
<thead>
<tr>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSTR</td>
<td>I</td>
<td>$MSTR = \frac{SSTR}{I-1}$</td>
<td>$E\sigma^2 + \sum_{i=1}^{I} \frac{n_i (\mu_i - \bar{\mu})^2}{I-1}$</td>
</tr>
<tr>
<td>SSE</td>
<td>$n-I$</td>
<td>$MSE = \frac{SSE}{n-I}$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>SSTO</td>
<td>$n-1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Therefore, we can construct the ANOVA table 1.

Comment: A lot of books or papers call SSTR and SSE as SSB or SSW respectively, where SSW means sum of squares of within groups and SSB means sum of squares of between groups.

Example. Consider the following table.

<table>
<thead>
<tr>
<th>Levels</th>
<th>Sample Size</th>
<th>Sample Mean</th>
<th>Sample Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_1$</td>
<td>$\bar{y}_1$</td>
<td>$S_1^2$</td>
</tr>
<tr>
<td>2</td>
<td>$n_2$</td>
<td>$\bar{y}_2$</td>
<td>$S_2^2$</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>I</td>
<td>$n_I$</td>
<td>$\bar{y}_I$</td>
<td>$S_I^2$</td>
</tr>
</tbody>
</table>

Then, we have $\bar{y}_. = \sum_{i=1}^{I} n_i \bar{y}_i / n$, $\hat{\sigma}^2 = \sum_{i=1}^{I} (n_i - 1) S_i^2 / (n - I)$, $E(\bar{y}_i) = \mu_i$, $E(\bar{y}_.) = \sum_{i=1}^{I} n_i \mu_i / n$. Therefore,

$$\bar{\mu} = \frac{\mu_1 + \cdots + \mu_I}{I}$$

if and only if $n_1 = \cdots = n_I$.

**F test for Equality of Factor Level Means.** Consider the test

$$H_0 : \mu_1 = \cdots = \mu_I. \quad \text{(2)}$$

Then, under $H_0$

$$F = \frac{MSTR}{MSE} \sim F_{I-1,n-I}. \quad \text{(1)}$$

Thus, $H_0$ is rejected if $F \geq F_{I-1,n-I}(1-\alpha)$.

**Alternative Form for One-Way ANOVA model.** Let $\mu = \mu$ and $\tau_i = \mu_i - \mu$. Then the one-way ANOVA model can be written as

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

4
where $\sum \tau_i = 0$. The book calls it *factor effects model*. Then, the LS estimates are

$$\hat{\tau}_i = \bar{Y}_i - \hat{\mu}$$

and

$$\hat{\mu} = \frac{1}{I} \sum_{i=1}^{I} \bar{Y}_i.$$ 

Testing 

$$H_0 : \tau_1 = \cdots = \tau_I = 0,$$

is based on the statistic in (1) and $H_0$ is rejected if $F \geq F_{I-1,n-I}(1 - \alpha)$.

*Matrix Expression* Let

$$(Y_{11} \quad Y_{12} \quad \cdots \quad Y_{1n_1} \quad Y_{21} \quad \cdots \quad Y_{2n_2} \quad \cdots \quad Y_{I-1,1} \quad \cdots \quad Y_{I-1,n_{I-1}} \quad Y_{I,1} \quad \cdots \quad Y_{I,n_I})$$

and

$$\begin{pmatrix}
\epsilon_{11} \\
\epsilon_{12} \\
\vdots  \\
\epsilon_{1n_1} \\
\epsilon_{21} \\
\vdots \\
\epsilon_{n_1n_1} \\
\epsilon_{I1} \\
\vdots  \\
\epsilon_{In_{I-1}} \\
\epsilon_{I1} \\
\vdots \\
\epsilon_{I,n_{I-1}} \\
\epsilon_{I1} \\
\vdots \\
\epsilon_{I1} \\
\epsilon_{I1} \\
\epsilon_{I1} \\
\epsilon_{I1} \\
\epsilon_{I1} \\
\epsilon_{I1}
\end{pmatrix}, \quad \beta = \begin{pmatrix}
\mu \\
\tau_1 \\
\vdots \\
\tau_{I-1}
\end{pmatrix},$$

and

$$X = \begin{pmatrix}
1_{n_1} & 1_{n_1} & \cdots & 0_{n_1} \\
1_{n_2} & 0_{n_2} & \cdots & 0_{n_2} \\
\vdots & \vdots & \ddots & \vdots \\
1_{n_{I-1}} & 0_{n_{I-1}} & \cdots & 1_{n_{I-1}} \\
1_{n_I} & -1_{n_I} & \cdots & -1_{n_I}
\end{pmatrix}.$$ 

Then, the matrix expression for factor effects model is

$$Y = X\beta + \epsilon.$$ 

**Chapter 17 Analysis of Factor Level Effects**

In the one way ANOVA model, it is very interesting to study the difference of the means. There are several options. The test based on (1) is a good test if $I < n$. However if $I = n$, that is $n_i = 1$ for all $i$, then the test based on (1) does not work any more. Normal Probability plot can tell us something about it.

**Normal Probability Plot to test when there is no replicate.** Suppose that $H_0 : \mu_1 = \cdots = \mu_I = \mu$ are correct. Then, the observations are

$$Y_i \sim N(\mu, \sigma^2).$$
Then,
\[ \hat{\mu} = \bar{Y} = \frac{1}{I} \sum_{i=1}^{I} Y_i \]
and
\[ \hat{\sigma}^2 = \frac{1}{I - 1} \sum_{i=1}^{I} (Y_i - \bar{Y})^2. \]

Let \( Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(I)} \) be the order statistic for the observations. Then, the density of \( (Y_{(1)}, \cdots, Y_{(I)}) \) is
\[ p(t_1, \cdots, t_I) = \frac{I!}{\sigma^I (2\pi)^{\frac{I}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{I} (t_i - \mu)^2 \right\} \]
if \( t_1 < t_2 < \cdots < t_I \) and 0 otherwise. Thus, the expectation of \( Y_{(i)} \) can be computed, which can be approximated by
\[ E(Y_{(i)}) = \mu + \sigma \Phi^{-1}\left( \frac{i - 0.375}{n + 0.25} \right). \]

Then the estimate is
\[ \hat{E}(Y_{(i)}) = \bar{Y} + \hat{\sigma} \Phi^{-1}\left( \frac{i - 0.375}{n + 0.25} \right). \]

Therefore, if the plot \((\bar{Y}_{(i)}, \hat{E}(Y_{(i)}))\) shows that the points are on the straight line
\[ \text{Vertical} = \bar{Y} + \hat{\sigma} \Phi^{-1}\left( \frac{i - 0.375}{n + 0.25} \right), \]
then we conclude that there is no significant difference between the means.

**Example.** Assume that \( I = n = 10 \) and the data are given the following table. There are two sets of data in this table. I use \( Y_{i1} \) and \( Y_{i2} \) to represent them. The graphs for normal probability plot are given in Figure 1. The first plot shows all the means are equal but the second not. The two straight lines have intercept 0 and slope 1.

**Comment:** Normal probability plot can give us a conclusion when there is no replicate in each cell. If there are replications, it requires the number of the replication be equal such that the variance of the average of the observation in each cell are equal.

**Comment:** The difference for normal probability plot and QQ-plot is the horizontal scale.

**A Robust Testing Method.** Assume that all \( n_i = 1 \). Then, there is no \( F \) test for \( H_0 \). Let \( \hat{\theta}_i = Y_i - \bar{Y} \), and let \( |\theta|_{(1)}, \cdots, |\theta|_{(I)} \) be the order statistic of \( |\theta_1|, \cdots, |\theta_I| \). Assume the medium of the order statistics is \( M \). Then,
\[ \frac{M}{\sigma} \approx 0.6745 \]
because 0.6745 is the median of the standard half normal. Thus, the estimate of \( \sigma \) is \( M/0.6745 = 1.483M \). Based on this fact, Lenth (1989) gave a procedure
Table 2: Detail for Normal Probability Plot

<table>
<thead>
<tr>
<th>$\Phi^{-1}(1 - 0.375)$</th>
<th>$Y_1^1$</th>
<th>$Y_1^1$</th>
<th>Expected(I)</th>
<th>$Y_2^2$</th>
<th>$Y_2^2$</th>
<th>Expected(II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.5466</td>
<td>0.6764</td>
<td>-1.7540</td>
<td>-1.6669</td>
<td>-0.9933</td>
<td>-0.9933</td>
<td>-0.6894</td>
</tr>
<tr>
<td>-1.0005</td>
<td>-1.7540</td>
<td>-1.2301</td>
<td>-1.1738</td>
<td>1.4451</td>
<td>-0.5024</td>
<td>0.1010</td>
</tr>
<tr>
<td>0.6554</td>
<td>-0.4150</td>
<td>-0.8748</td>
<td>-0.8622</td>
<td>-0.5024</td>
<td>0.4025</td>
<td>0.6004</td>
</tr>
<tr>
<td>-0.3755</td>
<td>-0.8748</td>
<td>-0.6317</td>
<td>-0.6093</td>
<td>1.8218</td>
<td>1.4451</td>
<td>1.0056</td>
</tr>
<tr>
<td>-0.1226</td>
<td>0.1832</td>
<td>-0.4150</td>
<td>-0.3810</td>
<td>0.4025</td>
<td>1.8218</td>
<td>1.3715</td>
</tr>
<tr>
<td>0.1226</td>
<td>-0.0847</td>
<td>-0.0847</td>
<td>-0.1596</td>
<td>2.1555</td>
<td>2.1555</td>
<td>1.7264</td>
</tr>
<tr>
<td>0.3755</td>
<td>0.1544</td>
<td>0.1544</td>
<td>0.0688</td>
<td>2.4029</td>
<td>2.4029</td>
<td>2.0923</td>
</tr>
<tr>
<td>0.6554</td>
<td>-1.2301</td>
<td>0.1832</td>
<td>0.3216</td>
<td>2.7466</td>
<td>2.7466</td>
<td>2.4975</td>
</tr>
<tr>
<td>1.0005</td>
<td>-0.6317</td>
<td>0.6764</td>
<td>0.6332</td>
<td>3.0570</td>
<td>2.9538</td>
<td>2.9969</td>
</tr>
<tr>
<td>1.5466</td>
<td>1.2733</td>
<td>1.2733</td>
<td>1.1264</td>
<td>2.9538</td>
<td>3.0570</td>
<td>3.7873</td>
</tr>
</tbody>
</table>

Figure 1: Normal Probability Plot
Table 3: Results for $t_{PSE,i}$

<table>
<thead>
<tr>
<th>$\theta_i^1$</th>
<th>$t_{PSE,i}^1$</th>
<th>$\theta_i^2$</th>
<th>$t_{PSE,i}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9467</td>
<td>1.2068</td>
<td>-2.5422</td>
<td>-2.1690</td>
</tr>
<tr>
<td>-1.4837</td>
<td>-1.8913</td>
<td>-0.1039</td>
<td>-0.0886</td>
</tr>
<tr>
<td>-0.1447</td>
<td>-0.1844</td>
<td>-2.0514</td>
<td>-1.7502</td>
</tr>
<tr>
<td>-0.6045</td>
<td>-0.7706</td>
<td>0.2728</td>
<td>0.2328</td>
</tr>
<tr>
<td>0.4535</td>
<td>0.5781</td>
<td>-1.1465</td>
<td>-0.9782</td>
</tr>
<tr>
<td>0.1856</td>
<td>0.2366</td>
<td>0.6066</td>
<td>0.5175</td>
</tr>
<tr>
<td>0.4247</td>
<td>0.5414</td>
<td>0.8540</td>
<td>0.7286</td>
</tr>
<tr>
<td>-0.9599</td>
<td>-1.2235</td>
<td>1.1976</td>
<td>1.0218</td>
</tr>
<tr>
<td>-0.3614</td>
<td>-0.4606</td>
<td>1.5081</td>
<td>1.2867</td>
</tr>
<tr>
<td>1.5436</td>
<td>1.9676</td>
<td>1.4049</td>
<td>1.1987</td>
</tr>
</tbody>
</table>

1. Let

$$PSE = 1.483 \times \text{median}_{\{ |\theta_i| < 2.57s_0 \}} |\theta_i|$$

where

$$s_0 = 1.483 \times \text{median} |\theta_i|.$$ 

2. Let

$$t_{PSE,i} = \frac{\theta_i}{PSE}.$$ 

Then, $\mu_i - \mu$ is declared significant if $t_{PSE,i}$ is greater than a critical value which derived from the standard normal distribution.

*Example:* Use the data in left Table 2. $\bar{Y} = -0.2703$. Then, $M = 0.5290$. Thus

$$s_0 = 1.483 \times 0.5290 = 0.7845.$$ 

There is no observation with absolute value greater than $2.57 \times 0.7845 = 2.0162$. Thus, $PSE = 1.483 \times 0.5290 = 0.7845$. Similarly, we have the right Table 2 with $PSE = 1.1721$. Therefore, we have table 3.

*Testing and Confidence Interval of Factor Level Effects.* The model in this section is based on mean cells model. *Confidence interval and testing for single mean.* It is clear that

$$\hat{\mu}_i = \bar{Y}_i \sim N(\mu_i, \frac{\sigma^2}{n_i}),$$

and

$$\hat{\sigma}^2 = MSE = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$
Thus,
\[ t^* = \frac{\bar{Y}_i - \mu_i}{std(Y_i)} = \sqrt{\frac{\bar{Y}_i - \mu_i}{\hat{\sigma}}} \sim t_{N-1}. \]

Therefore, the \((1 - \alpha)100\%\) confidence interval for \(\mu_i\) is
\[
[\bar{Y}_i - std(Y_i) t_{n-1}(1 - \alpha/2), \bar{Y}_i + std(Y_i) t_{n-1}(1 - \alpha/2)]
\]
\[=\left[\bar{Y}_i - \frac{\hat{\sigma}}{\sqrt{n_i}} t_{n-1}(1 - \alpha/2), \bar{Y}_i + \frac{\hat{\sigma}}{\sqrt{n_i}} t_{n-1}(1 - \alpha/2)\right].\]

To test \(H_0 : \mu_i = c\) if and only if the confidence interval contains the point \(c\).

Confidence interval and testing for difference of pair means. It is clear that \(\hat{\mu}_i\) and \(\hat{\mu}_{i'}\) are independent if \(i \neq i'\). Thus
\[
\hat{\mu}_i - \hat{\mu}_{i'} = \bar{Y}_i - \bar{Y}_{i'} \sim N(\mu_i - \mu_{i'}, \frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_{i'}}).
\]

Thus
\[
\frac{\bar{Y}_i - \bar{Y}_{i'}}{std(\bar{Y}_i - \bar{Y}_{i'})} = \frac{n_i + n_{i'}}{n_i n_{i'}} \frac{\bar{Y}_i - \bar{Y}_{i'}}{\hat{\sigma}} \sim t_{n-1}.
\]

Therefore, the \((1 - \alpha)100\%\) confidence interval for \(\mu_i - \mu_{i'}\) is
\[
[\bar{Y}_i - \bar{Y}_{i'} - std(\bar{Y}_i - \bar{Y}_{i'}) t_{N-1}(1 - \alpha/2), \bar{Y}_i - \bar{Y}_{i'} + std(\bar{Y}_i - \bar{Y}_{i'}) t_{N-1}(1 - \alpha/2)].
\]

If the confidence interval contains 0 then \(H_0 : \mu_i = \mu_{i'}\) is accepted; otherwise \(H_0\) is rejected.

Testing for contrast of factor means. The contrast of factor means is defined by
\[
L = \sum_{i=1}^{I} c_i \mu_i,
\]

where \(\sum c_i = 0\) and \(c_i\) are not all 0. Then the estimated of \(L\) is
\[
\hat{L} = \sum_{i=1}^{I} c_i \bar{Y}_i \sim N(L, \sum_{i=1}^{I} \frac{c_i^2 \sigma^2}{n_i}).
\]

Thus,
\[
std(\hat{L}) = \hat{\sigma} \sqrt{\sum_{i=1}^{I} \frac{c_i^2}{n_i}}.
\]

Thus, the \((1 - \alpha)100\%\) confidence interval is
\[
[\hat{L} - std(\hat{L}) t_{N-1}(1 - \alpha/2), \hat{L} + std(\hat{L}) t_{N-1}(1 - \alpha/2)].
\]

To test \(H_0 : L = c\) if and only if the confidence interval contains \(c\).
Comment: This testing results can be paralleled to factor effects model

\[ Y_{ij} = \mu + \tau_i + \epsilon_{ij} \]

where \( \sum_{i=1}^{I} \tau_i = 0 \). Only the confidence interval of the contrast of \( \tau_i \) can be computed.

Comment: For cell means model, the result does not require the contrast.

**Tukey Multiple Comparison Procedure.** Assume that \( Y_1, \cdots, Y_k \) and \( S^2 \) independently follow \( N(\mu, \sigma^2) \) and \( \sigma^2 \chi^2_{N} / N \) respectively. Then

\[ T = \frac{\max Y_i - \min Y_i}{S} \]

gives the *Tukey studentized range distribution* with parameter \( k \) and \( N \).

The *Tukey studentized range distribution* gives us the simultaneous confidence intervals for all pairs of \( \mu_i - \mu_i' \). The corresponding testing result gives us the simultaneous testing result for \( H_0 : \mu_i = \mu_i' \), all \( i \neq i' \) and the corresponding testing result.

Let us consider the mean cells model

\[ Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \cdots, I, \quad J = 1, \cdots, n_i, \]

where \( \epsilon_{ij} \sim N(0, \sigma^2) \). Let \( n = \sum n_i \),

\[ D_{i'i'} = \bar{Y}_i - \bar{Y}_{i'} \]

and

\[ \hat{\text{Var}}(D_{i'i'}) = \hat{\sigma}^2 \left( \frac{1}{n_i} + \frac{1}{n_{i'}} \right), \]

where

\[ \hat{\sigma}^2 = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2. \]

Then,

\[ T = \max \frac{D_{i'i'}}{\sqrt{\hat{\text{Var}}(D_{i'i'})}} > \frac{1}{\sqrt{2}} q_{I,n-I}(1 - \alpha) \]

is the result for testing \( H_0 : \mu_i = \mu_i' \) for \( 1 \leq i, i' \leq I \).

The confidence interval for \( \mu_i - \mu_i' \) is

\[ [D_{i'i'} - q_{I,n-I}(1 - \alpha)\sqrt{\hat{\text{Var}}(D_{i'i'})}/2, D_{i'i'} + q_{I,n-I}(1 - \alpha)\sqrt{\hat{\text{Var}}(D_{i'i'})}/2]. \]

If one of those confidence intervals for \( \mu_i - \mu_i' \) does not contain 0, then \( H_0 \) is rejected.

When all \( n_i \) are equal, the distribution of \( \sqrt{2T} \) follows *Tukey studentized range distribution* with degrees of freedom \( I \) and \( n - I \), denoted \( q_{I,n-I} \). When \( n_i \) are not all equal, Hochberg and Tamhane prove the type I error for such test is at most \( \alpha \). It is clear that the type I error is exactly \( \alpha \) when \( n_i \) are all equal.
Table 4: Data for Rust Inhibitor Example

<table>
<thead>
<tr>
<th>Index</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>43.9</td>
<td>89.8</td>
<td>68.4</td>
<td>36.2</td>
</tr>
<tr>
<td>2</td>
<td>39.0</td>
<td>87.1</td>
<td>69.3</td>
<td>45.2</td>
</tr>
<tr>
<td>3</td>
<td>46.7</td>
<td>92.7</td>
<td>68.5</td>
<td>40.7</td>
</tr>
<tr>
<td>4</td>
<td>43.8</td>
<td>90.6</td>
<td>66.4</td>
<td>40.5</td>
</tr>
<tr>
<td>5</td>
<td>44.2</td>
<td>87.7</td>
<td>70.0</td>
<td>39.3</td>
</tr>
<tr>
<td>6</td>
<td>47.7</td>
<td>92.4</td>
<td>68.1</td>
<td>40.3</td>
</tr>
<tr>
<td>7</td>
<td>43.6</td>
<td>86.1</td>
<td>70.6</td>
<td>43.2</td>
</tr>
<tr>
<td>8</td>
<td>38.9</td>
<td>88.1</td>
<td>65.2</td>
<td>38.7</td>
</tr>
<tr>
<td>9</td>
<td>43.6</td>
<td>90.8</td>
<td>63.8</td>
<td>40.9</td>
</tr>
<tr>
<td>10</td>
<td>40.0</td>
<td>89.1</td>
<td>69.2</td>
<td>39.7</td>
</tr>
</tbody>
</table>

\[ \bar{Y}_i \times 43.14 \quad 89.44 \quad 67.95 \quad 40.47 \]

Table 5: Difference and Confidence Interval

<table>
<thead>
<tr>
<th></th>
<th>( \mu_1 - \mu_2 )</th>
<th>( \mu_1 - \mu_3 )</th>
<th>( \mu_1 - \mu_4 )</th>
<th>( \mu_2 - \mu_3 )</th>
<th>( \mu_2 - \mu_4 )</th>
<th>( \mu_3 - \mu_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-46.30</td>
<td>-24.81</td>
<td>2.67</td>
<td>21.49</td>
<td>48.97</td>
<td>27.48</td>
</tr>
<tr>
<td>CI</td>
<td>(-49.29, -43.31)</td>
<td>(-27.80, -21.82)</td>
<td>(-0.32, 5.66)</td>
<td>(18.50, 24.48)</td>
<td>(45.98, 51.96)</td>
<td>(24.49, 30.47)</td>
</tr>
</tbody>
</table>

Comment: This method is known to be generally the most effective among conservative methods for the one-way ANOVA model, that is the type II error is the smallest.

Example: Let us use the dataset of Table 17.2, textbook, where \( \hat{\sigma}^2 = 0.1406 \). Thus, \( \hat{\sigma}^2(1/10 + 1/10) = 1.23 \) for all pairs. Here \( q_{4,36}(0.95) = 3.814 \). Thus, we have

\[
q_{t,n-1}(1-\alpha)\sqrt{\text{Var}(D_{ii'})}/2 = 2.991
\]

for all pairs. The difference of the average are given in Table 5, where \( D_{ii'} = \bar{Y}_i - \bar{Y}_{i'} \). The 95% confidence interval are also given in this table.

Scheffe Multiple Comparison Procedure. gives the method for simultaneous confidence interval for infinite many contrasts of the means, which is defined by

\[
L = \sum_{i=1}^{l} c_i \mu_i,
\]

where \( \sum c_i = 0 \) and \( c_i \) are not all zero. It is clear that the dimension of the contrast space
is $I - 1$, and the estimated of the contrast is

$$\hat{L} = \sum_{i=1}^{I} c_i \bar{Y}_i.$$ 

Thus, the result for the confidence interval is

$$[\hat{L} - S \sqrt{\text{MSE} \sum c_i^2 / n_i}, \hat{L} + S \sqrt{\text{MSE} \sum c_i^2 / n_i}].$$

where

$$S^2 = (I - 1)F_{I-1,n-I}(1 - \alpha).$$

The simultaneous test for $H_0 : L = 0$ rejects $H_0$ if

$$F^* = \frac{\hat{L}^2}{(I - 1)\text{Var}(\hat{L})}$$

is greater than the critical value $(3F_{I-1,n-I}(1 - \alpha))$.

If one of these infinite confidence intervals does not contain 0, then $H_0$ is rejected.

**Example.** Let us using the data in Table 4. Then, $I = 4, n = 40, \text{MSE} = 6.14$. Therefore

$$S^2 = 3F_{3,36}(1 - 0.05) = 8.5988.$$

Consider $L = \mu_1 + \mu_2 - \mu_3 - \mu_4$. Then

$$\hat{L} = 43.14 + 89.44 - 67.95 - 40.47 = 24.16,$$

and

$$\text{MSE} \sum c_i^2 / n_i = 6.14 \times \frac{4}{10} = 2.456.$$ 

Then, the Scheffe CI for $\sum c_i \mu_i$ is

$$[24.16 - \sqrt{8.5988 \times 2.456}, 24.16 + \sqrt{8.5988 \times 2.456}] = [19.56, 28.76].$$

**Comment:** If the $F$ test shows $H_0$ is not correct, then the corresponding scheffe multiple comparison procedure should find at least one contrast is significant different from zero.

**Bonferroni Multiple Comparison Procedure.** The result for Bonferroni multiple comparison procedure for $k$ contrasts $L$ is

$$[L - B \sqrt{\text{MSE} \sum c_i^2 / n_i}, L + B \sqrt{\text{MSE} \sum c_i^2 / n_i}],$$

where

$$B = t_{n-l}(1 - \alpha/2k).$$
Table 6: Holm Testing Result

<table>
<thead>
<tr>
<th>$\hat{L}_i$</th>
<th>$s(L_i)$</th>
<th>$t_i$</th>
<th>$p$-value</th>
<th>critical</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>-9.35</td>
<td>1.5</td>
<td>-6.23</td>
<td>0.000016</td>
<td>0.01</td>
<td>$H_a$</td>
</tr>
<tr>
<td>-3.25</td>
<td>1.5</td>
<td>-2.17</td>
<td>0.0465</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.20</td>
<td>2.05</td>
<td>0.585</td>
<td>0.567</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-4.90</td>
<td>2.18</td>
<td>-2.25</td>
<td>0.0399</td>
<td>0.0167</td>
<td>$H_0$</td>
</tr>
<tr>
<td>-13.8</td>
<td>2.05</td>
<td>-6.73</td>
<td>0.000007</td>
<td>0.0083</td>
<td>$H_a$</td>
</tr>
<tr>
<td>-7.70</td>
<td>2.18</td>
<td>-3.53</td>
<td>0.00303</td>
<td>0.0125</td>
<td>$H_a$</td>
</tr>
</tbody>
</table>

The simultaneous test for $H_0 : L = 0$ for one of those $k$ independent contrasts $L$ rejects $H_0$ if none of such intervals contains 0.

**Holm Simultaneous Testing Procedure.** Holm simultaneous testing problem is considering the simultaneous test for

$$H_0 : L_k = 0, \ k = 1, \ldots, g,$$

where $L_k = \sum_{i=1}^{I} c_{i,k} \mu_i$ is a contrast for each $k$. Bonferroni testing result is conservative. Holm method is a refined Bonferroni method. It wants to test whether the remains are different if the significant is excluded. It has many conclusions for the testing result.

Assume that the studentized value is derived as

$$t_k = \frac{\sum_{i=1}^{I} c_{i,k} \bar{Y}_i}{\hat{\sigma} \sqrt{\sum_{i=1}^{I} c_{i,k}^2 / n_i}}.$$

Assume that $t_k$ is sorted by $|t_1| > |t_2| > \cdots > |t_g|$. Then the procedure is

1. If $|t_1| \leq t_{n-f}(1 - \alpha/2g)$, then $H_0$ is concluded and the procedure stops; otherwise goes to step 2.

2. If $|t_k| \leq t_{n-f}(1 - \alpha/2(g - k + 1))$, then $H_0$ is conclude and the procedure stops; otherwise let $k = k + 1$ check the next one.

3. $H_0$ is rejected if no $H_0$ is concluded in step 2.

**Example:** This example is based on the dataset given in Table 16.1 textbook. There are 4 different designs and 5 observations in design 1, 2, 3 and 4 observation in design 4. The means are $Y_1 = 14.6, Y_2 = 13.4, Y_3 = 19.5, Y_4 = 27.2$. Let us choose

$$L_1 = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}, \quad L_2 = \frac{\mu_1 + \mu_3}{2} - \frac{\mu_2 + \mu_4}{2},$$

$$L_3 = \mu_1 - \mu_2, \quad L_4 = \mu_1 - \mu_3, \quad L_5 = \mu_2 - \mu_4, \quad L_6 = \mu_3 - \mu_4.$$
Then, $H_0$ is concluded.

**ANOVA model for Quantitative Factor.** Let us consider the lack of fit full model,

$$Y_{ij} = \mu + x_i \beta_i + \epsilon_{ij},$$

where $j = 1, \cdots, n_i$, $i = 1, \cdots, I$ and $n = \sum n_i$, $\epsilon_{ij} \sim N(0, \sigma^2)$. The estimate of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2.$$

Thus,

$$SSLF = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y})^2.$$

Then, consider the lack of fit reduced model

$$Y_{ij} = \mu + x_i \beta + \epsilon_{ij}.$$

Then,

$$SSR = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{I} n_i (\bar{Y}_i - \mu - x_i \hat{\beta})^2.$$

Then, the estimate of $\sigma^2$ based on reduced model is

$$\hat{\sigma}^2_R = \frac{SSR}{n - 2}.$$

There are two ways to fit the linear model from this data. The first one fits the model for all the observations by unweighted least squared method, and the second fits the model

$$\bar{Y}_i = \mu + x_i \beta + \bar{\epsilon}_i,$$

by weighted least squared method. Then, the estimate of $\sigma^2$ based on the first method is

$$\hat{\sigma}^2_1 = \hat{\sigma}^2_R,$$

and the estimate of $\sigma^2$ based on the second method is

$$\hat{\sigma}^2_2 = \sum_{i=1}^{I} n_i (\bar{Y}_i - \mu - x_i \hat{\beta})^2 / (I - 2).$$

**Comment:** If lack of fit is huge, then $\hat{\sigma}^2_1 \approx \hat{n} \hat{\sigma}^2_2$; otherwise $\hat{\sigma}^2_1 \approx \hat{\sigma}^2_2$. The estimates of $\mu$ and $\beta$ are the same in these methods. However, if the lack of fit is huge, the variance of the two methods for $\beta$ is

$$\frac{I}{\hat{n}} Var(\hat{\beta}^2) \approx Var(\hat{\beta}^1) \sum_{i=1}^{I} \frac{1}{n_i},$$

if the lack of fit is huge; otherwise,

$$Var(\hat{\beta}^2) \approx Var(\hat{\beta}^1).$$
Chapter 18: ANOVA Diagnostics and Remedial Measures

Residual analysis and plots. All the assumptions should be diagnosed here. There are:

1. Outliers- jackknife residual.
3. Nonindependency of error variance- plot $\hat{\epsilon}_k$ versus $\hat{\epsilon}_k$.
5. Partial residual plot for checking important explanatory variables.

Tests for Constant of Variances. There are several numeric tests for

$$H_0 : \sigma_1^2 = \cdots = \sigma_I^2.$$ 

Hartley Test. The UMVU estimator for each $i$ is

$$S_i^2 = \frac{1}{J-1} \sum_{i=1}^{J} (Y_{ij} - \bar{Y}_i)^2.$$ 

Assume that $H_0$ is true and all $n_i$ are equal to $J$. Then

$$H = \frac{\max S_i^2}{\min S_i^2}$$

follows Hartley distribution with parameter $I$ and $J-1$. The critical values for Hartley distribution are given in Table B.10, textbook. Therefore, $H_0$ is rejected if

$$H_0 > H_{I,J-1}(1-\alpha).$$

This test highly depends on equal size in each cell.

Modified Levene Test. This is also called Brown-Forsythe test. Let $\bar{Y}_i = \text{median}\{Y_{ij}\}$ and

$$d_{ij} = |Y_{ij} - \bar{Y}_i|.$$ 

Let $\bar{d}_i = \sum_{j=1}^{n_i} d_{ij}/n_i$ and $\bar{d} = \sum \bar{d}_i/n$, and

$$NUM = \frac{\sum_{i=1}^{I} n_i (\bar{d}_i - \bar{d})^2}{I-1}$$

and

$$DEN = \frac{\sum \sum (d_{ij} - \bar{d}_i)^2}{n-I}.$$
<table>
<thead>
<tr>
<th></th>
<th>$Y_{1j}$</th>
<th>$Y_{2j}$</th>
<th>$Y_{3j}$</th>
<th>$Y_{4j}$</th>
<th>$Y_{5j}$</th>
<th>$d_{1j}$</th>
<th>$d_{2j}$</th>
<th>$d_{3j}$</th>
<th>$d_{4j}$</th>
<th>$d_{5j}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>14.87</td>
<td>18.43</td>
<td>16.95</td>
<td>8.59</td>
<td>11.55</td>
<td>0.300</td>
<td>0.165</td>
<td>1.695</td>
<td>1.420</td>
<td>0.555</td>
</tr>
<tr>
<td>16.81</td>
<td>18.76</td>
<td>12.28</td>
<td>10.90</td>
<td>13.36</td>
<td></td>
<td>1.640</td>
<td>0.165</td>
<td>2.975</td>
<td>0.890</td>
<td>1.255</td>
</tr>
<tr>
<td>15.83</td>
<td>20.12</td>
<td>12.00</td>
<td>8.60</td>
<td>13.64</td>
<td></td>
<td>0.660</td>
<td>1.525</td>
<td>3.255</td>
<td>1.410</td>
<td>1.535</td>
</tr>
<tr>
<td>15.47</td>
<td>19.11</td>
<td>13.18</td>
<td>10.13</td>
<td>12.16</td>
<td></td>
<td>0.300</td>
<td>0.515</td>
<td>2.075</td>
<td>0.120</td>
<td>0.055</td>
</tr>
<tr>
<td>13.60</td>
<td>19.81</td>
<td>14.99</td>
<td>10.28</td>
<td>11.62</td>
<td></td>
<td>1.570</td>
<td>1.215</td>
<td>0.265</td>
<td>0.270</td>
<td>0.485</td>
</tr>
<tr>
<td>14.76</td>
<td>18.43</td>
<td>15.76</td>
<td>9.98</td>
<td>12.39</td>
<td></td>
<td>0.410</td>
<td>0.165</td>
<td>0.505</td>
<td>0.030</td>
<td>0.285</td>
</tr>
<tr>
<td>17.40</td>
<td>17.16</td>
<td>19.35</td>
<td>9.41</td>
<td>12.05</td>
<td></td>
<td>2.230</td>
<td>1.435</td>
<td>4.095</td>
<td>0.600</td>
<td>0.055</td>
</tr>
<tr>
<td>14.62</td>
<td>16.40</td>
<td>15.52</td>
<td>10.04</td>
<td>11.95</td>
<td></td>
<td>0.550</td>
<td>2.195</td>
<td>0.265</td>
<td>0.030</td>
<td>0.155</td>
</tr>
<tr>
<td>Mean</td>
<td>15.420</td>
<td>18.527</td>
<td>15.004</td>
<td>9.741</td>
<td>12.340</td>
<td>0.9575</td>
<td>0.9225</td>
<td>1.8913</td>
<td>0.5962</td>
<td>0.5475</td>
</tr>
<tr>
<td>Median</td>
<td>15.170</td>
<td>18.595</td>
<td>15.255</td>
<td>10.010</td>
<td>12.105</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance</td>
<td>1.531</td>
<td>1.570</td>
<td>6.183</td>
<td>0.667</td>
<td>0.592</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then, under $H_0$,

$$F_L^* = \frac{\text{NUM}}{\text{DEN}}$$

approximately follows $F$-distribution with parameter $I - 1$ and $N - I$. Thus $H_0$ is rejected if $F_L^* \geq F_{I-1,N-I}(1 - \alpha)$.

**Example:** Table 7 is based on the dataset on Page 765, textbook.

$$H = \max \frac{\text{max} s_i^2}{\text{min} s_i^2} = \frac{6.183}{0.592} = 10.444 > H(0.95, 5, 7) = 9.70.$$  

Thus, $H_0 : \sigma_1^2 = \cdots = \sigma_5^2$ is rejected.

It is clear that for Modified Levene Test, we have $MSTR = 2.3372$ and $MSE = 0.7960$. Thus, $F_L^* = 2.3372/0.7960 = 2.936 > F_{4,35}(1 - 0.05) = 2.64$. Thus $H_0$ is rejected.

**Methods for Unequal Variances.**

**Weighted Least Square Method.** Consider the cell means model

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

where $\epsilon \sim N(0, \sigma_i^2)$. If the weight is unknown, then the estimate of $\sigma_i^2$ is

$$\hat{\sigma}_i^2 = s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2.$$  

Thus, the estimate of the weight is

$$w_i = \frac{1}{s_i^2}.$$
and the model can be fitted by weighted least squared method. The test for

\[ H_0 : \mu_1 = \cdots = \mu_I \]

can also be derived.

*Transformation.* Box-Cox transformation is another option. The option requires the response to be positive.

**Nonparametric Rank F test.**

Let \( Y_{ij} \) be ranked as \( Y_{(k)} \), \( k = 1, \cdots, n \), that is \( Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)} \). Let \( R_{ij} \) be the rank of \( Y_{ij} \). Let

\[ F_R^* = \frac{MSTR}{MSE}, \]

where

\[ MSTR = \frac{1}{I-1} \sum_{i=1}^{I} n_i (\bar{R}_i - \bar{R}_\cdot)^2 \]

\[ MSE = \frac{1}{n-I} \sum \sum (R_{ij} - \bar{R}_\cdot)^2, \]

and

\[ \bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij} \]

and

\[ \bar{R}_\cdot = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} R_{ij} = \frac{n+1}{2}. \]

Then, \( H_0 : \mu_1 = \cdots = \mu_I \) is rejected if \( F_R^* \) is large. It can be proved that under \( H_0 \), \( F_R^* \) follows an \( F \)-distribution with \( df_1 = I-1 \) and \( df_2 = n-I \). Thus, \( H_0 \) is rejected if

\[ F_R^* > F_{I-1,n-I}(1-\alpha). \]

*Comment:* The result is based on the new dataset based on \( R_{ij} \) which generated from the rank of the original dataset with assumption

\[ R_{ij} = \beta_i + \epsilon_{ij}, \]

where \( \epsilon_{ij} \sim N(0, \sigma^2) \).

*Comment:* An equivalent test, called *Kruskal-Wallis* test is consider

\[ X_{KW}^2 = (n-1) \frac{SSTR}{SSTO}, \]

and under \( H_0 \), \( X_{KW}^2 \) is approximated by \( \chi^2_{I-1} \). \( H_0 \) is rejected for large values.
If the rank $F$ test shows that $\mu_i$ are not all equal, then a further pairwise test for the difference of the cell mean is required. A test is based on the results for the confidence intervals for the difference of $E(R_i - R_{i'})$. The result of the CI is based on the fact that

$$\bar{R}_i \sim \text{approx } N(E(R_i), \frac{n(n+1)}{12n_i}).$$

Thus, the CI for $E(R_i - R_{i'})$ is

$$\bar{R}_i - \bar{R}_{i'} \pm C \left[ \frac{n(n+1)}{12} \left( \frac{1}{n_i} + \frac{1}{n'_i} \right) \right]^{\frac{1}{2}},$$

where $C$ can be chosen by some rules. One option is the Bonferrini rule, that is

$$B = z(1 - \frac{\alpha}{2g}), \quad g = \frac{I(I-1)}{2}.$$

Comment (not required): The estimate of the variance is based on the following fact. Assume that $H_0$ is true. Then for a particular $i$, it is equivalent to pickup $n_i$ numbers from $1, \cdots, n$. Thus, there are $\frac{n!}{(n-n_i)!n_i!}$ options and the points is to compute

$$RTS_i = \frac{1}{n_i} [R_{i1}^2 + R_{i2}^2 + \cdots + R_{in_i}^2].$$

There is no preference among $1, \cdots, n$. Thus, the expectation is

$$E(RTS_i) = \frac{1}{n} \sum_{k=1}^{n} k^2 = \frac{(n + 1)(2n + 1)}{6}.$$
Chapter 19: Two-Way ANOVA model

Assume that the study concerns the response related to two factor variables. Then, the model is called two-way ANOVA model or two-factor ANOVA model.

**Model and Matrix Expression.** Assume that factor $A$ has $I$ levels and factor $B$ has $J$ levels. In each level combination of factor $A$ and $B$, there are $n_{ij} \geq 1$ replicate observations. Then, the mean cells model is

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

where $\epsilon_{ijk} \sim i.i.d \ N(0, \sigma^2)$, $i = 1, \cdots, I$ and $j = 1, \cdots, J, k = 1, \cdots, n_{ij}, n_{ij} \geq 1$. Let

$$n_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}, \ n_{+i} = \sum_{j=1}^{J} n_{ij}, \ n_{i+} = \sum_{i=1}^{I} n_{ij}.$$  

Let us define the following matrix

$$Y = \begin{pmatrix} Y_{111} \\ Y_{112} \\ \vdots \\ Y_{IJn_{IJ}} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{IJ} \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_{111} \\ \epsilon_{112} \\ \vdots \\ \epsilon_{IJn_{IJ}} \end{pmatrix},$$

and

$$X = \begin{pmatrix} 1_{n_{11}} & 0_{n_{11}} & \cdots & 0_{n_{11}} \\ 0_{n_{12}} & 1_{n_{12}} & \cdots & 0_{n_{12}} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_{IJ}} & 0_{n_{IJ}} & \cdots & 1_{n_{IJ}} \end{pmatrix}.$$  

Then, the matrix expression of the model is

$$Y = X\beta + \epsilon.$$  

Then, we have

$$\hat{\beta} = (X^tX)^{-1}X^tY.$$  

Clearly

$$X^tX = \begin{pmatrix} n_{11} & 0 & \cdots & 0 \\ 0 & n_{12} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_{IJ} \end{pmatrix}.$$  

Thus, the LS estimates of the parameter is

$$\hat{\mu}_{ij} = \bar{Y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} Y_{ijk} \sim N(\mu_{ij}, \frac{\sigma^2}{n_{ij}}),$$
and

\[ \hat{\sigma}^2 = \frac{1}{n_{++} - IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij})^2. \]

Further, let us denote

\[ \bar{Y}_{i.} = \frac{1}{n_{i+}} \sum_{j,k} Y_{ijk}, \quad \bar{Y}_{.j} = \frac{1}{n_{+j}} \sum_{i,k} Y_{ijk}, \quad \bar{Y}_{..} = \frac{1}{n} \sum_{i,j,k} Y_{ijk}. \]

**Main Effect and Interaction Effect.** Let us look at the factor effect model

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad (2) \]

where \( i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad k = 1, \cdots, n_{ij} \), and

\[ \sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0. \]

Let us denote

\[ \mu_i = \frac{1}{J} \sum_{j=1}^{J} \mu_{ij} \]

and

\[ \mu_{.j} = \frac{1}{I} \sum_{i=1}^{I} \mu_{ij}. \]

Then, it can be proved that the relation between the parameters in the mean cells model and the factor effect model is

\[ \mu = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij} \]

and

\[ \alpha_i = \mu_i - \mu, \quad \beta_j = \mu_{.j} - \mu, \]

and

\[ (\alpha\beta)_{ij} = \mu_{ij} - \alpha_i - \beta_j - \mu = \mu_{ij} - \mu_i - \mu_{.j} + \mu. \]

Thus, the estimated of the parameters for the factor effect model is given in Table 8. The variances of those estimates are complicated and thus not in the table.

**Explanation of the parameters.** The parameter \( \alpha_i \) is called the main effect of factor \( A \) at level \( i \); \( \beta_j \) is called the main effect of factor \( B \) at level \( j \); and \( (\alpha\beta)_{ij} \) is called the two-factor interaction effect of factor \( A \) and \( B \) at level \( i \) and \( j \).

If \( I = J = 2 \), then \( \alpha_1 = -\alpha_2, \quad \beta_1 = -\beta_2, \quad (\alpha\beta)_{11} = (\alpha\beta)_{22} = -(\alpha\beta)_{12} = -(\alpha\beta)_{21} \). Then, we called them main effect of \( A \), main effect of \( B \), and interaction effect of \( A \) and \( B \).

**Example:** Suppose that \( I = J = 2 \) and \( \mu_{11} = 6, \quad \mu_{12} = 10, \quad \mu_{21} = 8, \quad \mu_{22} = 4 \). Then, we have \( \mu = 7, \quad \alpha_1 = -\alpha_2 = 1, \quad \beta_1 = -\beta_2 = 0, \quad (\alpha\beta)_{11} = -(\alpha\beta)_{12} = -(\alpha\beta)_{21} = (\alpha\beta)_{22} = -2. \)
Table 8: Estimate of the Parameters for Factor Effect Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>$\hat{\mu} = \frac{1}{IJ} \sum_i \sum_j \bar{Y}_{ij}$</td>
</tr>
<tr>
<td>$\alpha_i$</td>
<td>$\hat{\alpha}<em>i = \frac{1}{J} \sum_j \bar{Y}</em>{ij} - \hat{\mu}$</td>
</tr>
<tr>
<td>$\beta_j$</td>
<td>$\hat{\beta}<em>j = \frac{1}{I} \sum_i \bar{Y}</em>{ij} - \hat{\mu}$</td>
</tr>
<tr>
<td>$(\alpha\beta)_{ij}$</td>
<td>$(\hat{\alpha}\hat{\beta})<em>{ij} = \bar{Y}</em>{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu}$</td>
</tr>
</tbody>
</table>

Table 9: Dataset on Page 804, Textbook

<table>
<thead>
<tr>
<th></th>
<th>$j = 1$</th>
<th>$j = 2$</th>
<th>$j = 3$</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i = 1$</td>
<td>$\hat{\mu}_{11} = 11$</td>
<td>$\hat{\mu}_{12} = 13$</td>
<td>$\hat{\mu}_{13} = 18$</td>
<td>14</td>
</tr>
<tr>
<td>$i = 2$</td>
<td>$\hat{\mu}_{21} = 7$</td>
<td>$\hat{\mu}_{22} = 9$</td>
<td>$\hat{\mu}_{23} = 14$</td>
<td>10</td>
</tr>
<tr>
<td>Average</td>
<td>9</td>
<td>11</td>
<td>16</td>
<td>12</td>
</tr>
</tbody>
</table>

**Example:** Let us use the dataset on Page 804, textbook. There are two levels in Factor $A$ and three levels in factor $B$. The data is given in Table 9. Then, we have $\hat{\mu} = 12$, $\hat{\alpha}_1 = 14 - 12 = 2$, $\hat{\beta}_1 = 9 - 12 = -3$, $\hat{\beta}_2 = 11 - 12 = -1$, $(\hat{\alpha}\hat{\beta})_{ij} = 0$, for all $i$ and $j$.

**Equal Size Case (Balanced Case).** Balanced case means that $n_{11} = n_{12} = \cdots = n_{IJ} = K$. Let us study the total sum of squares

$$SSTO = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{..})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} (Y_{ijk} - \bar{Y}_{ij} + \bar{Y}_{ij} - \bar{Y}_{..})^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{ij}} [(\bar{Y}_{ijk} - \bar{Y}_{ij})^2 + (\bar{Y}_{ij} - \bar{Y}_{..})^2]$$

$$= SSE + \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}(\bar{Y}_{ij} - \bar{Y}_{..})^2$$

$$= SSE + SSTR.$$ 

It is clear that

$$SSTR = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}(\bar{Y}_{ij} - \bar{Y}_{..})^2$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}[(\bar{Y}_{i..} - \bar{Y}_{..})^2 + (\bar{Y}_{ij} - \bar{Y}_{i..})^2]$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}[(\bar{Y}_{i..} - \bar{Y}_{..})^2 + (\bar{Y}_{..} - \bar{Y}_{..})^2 + (\bar{Y}_{ij} - \bar{Y}_{ij} - \bar{Y}_{i..} + \bar{Y}_{i..})^2].$$
Table 10: ANOVA Table for Two-Way Balanced Model

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$KJ \sum (\bar{Y}<em>{i..} - \bar{Y}</em>{...})^2$</td>
<td>$I - 1$</td>
<td>$\frac{SSA}{(I-1)}$</td>
<td>$\sigma^2 + JK \sum_{i=1}^{I} \alpha_i^2$</td>
</tr>
<tr>
<td>B</td>
<td>$KI \sum_j (\bar{Y}<em>{.j} - \bar{Y}</em>{...})^2$</td>
<td>$J - 1$</td>
<td>$\frac{SSB}{(J-1)}$</td>
<td>$\sigma^2 + IK \sum_{j=1}^{J} \beta_j^2$</td>
</tr>
<tr>
<td>AB</td>
<td>$K \sum_i \sum_j (\bar{Y}<em>{ij} - \bar{Y}</em>{.j} - \bar{Y}<em>{i..} + \bar{Y}</em>{...})^2$</td>
<td>$(I - 1)(J - 1)$</td>
<td>$\frac{SSAB}{(I-1)(J-1)}$</td>
<td>$\sigma^2 + K \sum_{i=1}^{I} \sum_{j=1}^{J} (\alpha\beta)_{ij}^2$</td>
</tr>
<tr>
<td>Error</td>
<td>$\sum_i \sum_j \sum_k [(Y_{ijk} - \hat{Y}_{ijk})^2]$</td>
<td>$IJ(K - 1)$</td>
<td>$\frac{SSE}{IJ(K-1)}$</td>
<td>$\sigma^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2$</td>
<td>$IJK - 1$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When $n_{ij} = K$ for all $i, j$ with the zero-sum constraint, then we have

$$SSA = KJ \sum_{i=1}^{I} (\bar{Y}_{i..} - \bar{Y}_{...})^2 = KJ \sum_{i=1}^{I} \hat{\alpha}_i^2$$

$$SSB = KI \sum_{j=1}^{J} (\bar{Y}_{.j} - \bar{Y}_{...})^2 = KI \sum_{j=1}^{J} \hat{\beta}_j^2$$

and

$$SSAB = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\bar{Y}_{ij} - \bar{Y}_{.j} - \bar{Y}_{i..} + \bar{Y}_{...})^2 = K \sum_{i=1}^{I} \sum_{j=1}^{J} (\hat{\alpha}\hat{\beta})_{ij}^2.$$ Thus, we have Table 10.

**Comment:** It can be proved that $SSA$, $SSB$, and $SSAB$ are independent for the balanced case.

**Comment:** If $n_{ij}$ are not all equal, $\hat{\alpha}_i \neq \bar{Y}_{i..} - \bar{Y}_{...}$, thus, the last steps for the three equations are not correct.

**Test of Effects.** There are three hypotheses for two-way ANOVA model. Test for main effect of factor A,

$$H_0: \alpha_1 = \cdots = \alpha_I = 0 \leftrightarrow H_A: \alpha_1, \cdots, \alpha_I, \text{ not all equal;}$$

test for main effect of factor B,

$$H_0: \beta_1 = \cdots = \beta_J = 0 \leftrightarrow H_A: \beta_1, \cdots, \beta_J, \text{ not all equal;}$$

Test for two-factor interaction effect of A and B,

$$H_0: (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{IJ} = 0 \leftrightarrow H_A: (\alpha\beta)_{11}, \cdots, (\alpha\beta)_{IJ}, \text{ not all equal.}$$

Based on the balanced case and let

$$F_A = \frac{MSA}{MSE}, \quad F_B = \frac{MSB}{MSE}, \quad F_{AB} = \frac{MSAB}{MSE},$$

22
then the null of first test is rejected if $F_A > F_{I-1,n-IJ}(1-\alpha)$; the null of the second is rejected if $F_B > F_{J-1,n-IJ}(1-\alpha)$; and the null of the third is rejected if $F_{AB} > F_{(I-1)(J-1),n-IJ}(1-\alpha)$.

**Example:** The example is based on data on Page 833, textbook. There are two factors and one response. Factor $A$ has three levels and factor $B$ has two levels. There are 2 replicates within each combination. Thus $n_{ij} = 2$ for all $i$ and $j$. The summary of the results are in Table 11, where $\hat{\mu}_{11} = 45$, $\hat{\mu}_{12} = 43$, $\hat{\mu}_{21} = 65$, $\hat{\mu}_{22} = 69$, $\hat{\mu}_{31} = 40$, $\hat{\mu}_{12} = 44$, and $\hat{\sigma} = 3.21455$.

### Case When Data Not Replicated

Assume that $n_{ij} = 1$ for all $n_{ij}$. Then, the model becomes

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ij},$$

where the constrain is still zero-sum. Then, we have

$$\hat{\alpha}_i = \bar{Y}_i - \bar{Y}, \quad \hat{\beta}_j = \bar{Y}_j - \bar{Y}.$$  

and

$$(\hat{\alpha}\hat{\beta})_{ij} = Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}.$$

The first method to check $H_0$ is the normal probability plot. If the result shows that interaction effect is not important (significant at a relative high level), then an $F$ table can still be constructed.

\begin{table}[h]
\centering
\caption{Summary of the result}
\begin{tabular}{lcccc}
\hline
Parameter & Estimate & Std & P-value \\
\hline
$\mu$ & 51 & 0.92796 & < .0001 \\
$\alpha_1$ & -1 & 0.92796 & 0.3226 \\
$\alpha_2$ & 1 & 0.92796 & 0.3226 \\
$\beta_1$ & -7 & 1.31233 & 0.0018 \\
$\beta_2$ & 16 & 1.31233 & < 0.0001 \\
$\beta_3$ & -9 & 1.31233 & 0.0005 \\
$(\alpha\beta)_{11}$ & 2 & 1.31233 & 0.1783 \\
$(\alpha\beta)_{12}$ & -1 & 1.31233 & 0.4749 \\
$(\alpha\beta)_{13}$ & -1 & 1.31233 & 0.4749 \\
$(\alpha\beta)_{21}$ & -2 & 1.31233 & 0.1783 \\
$(\alpha\beta)_{22}$ & 1 & 1.31233 & 0.4749 \\
$(\alpha\beta)_{23}$ & 1 & 1.31233 & 0.4749 \\
\hline
\end{tabular}
\end{table}
Based on Table 10, let
\[ F_A^* = \frac{MSA}{MSAB}, \quad F_B^* = \frac{MSB}{MSAB}. \]

Then, the test for main effect of \( A \) rejects \( H_0 \) if \( F_A^* > F_{I-1,(I-1)(J-1)}(1-\alpha) \), and the test for main effect of \( B \) rejects \( H_0 \) if \( F_B^* > F_{J-1,(I-1)(J-1)}(1-\alpha) \).

**Comment:** If the interaction effect does not exist, the model is called additive model. If the model is additive and \( n_{ij} = K > 1 \) for all \( i \) and \( j \), then let
\[ SSE^* = SSE + SSAB \]
and
\[ MSE^* = \frac{SSE + SSAB}{IJK - I - J + 1}. \]

Let
\[ F_A^* = \frac{MSA}{MSE^*} \]
and
\[ F_B^* = \frac{MSB}{MSE^*}. \]

Then, the tests based on \( F_A^* \) and \( F_B^* \) is more powerful than the tests based on \( F_A \) and \( F_B \).

**Tukey test for additivity.** If there is no replication in each cell, then the general interaction model can not be solved. The full model in this test is assumed the interaction effect is the quadratic effect with same coefficient. The full model is
\[ Y_{ij} = \mu + \alpha_i + \beta_j + D\alpha_i\beta_j + e_{ij}. \]

Thus, the parameter space is an \((IJ+1)\)-dimensional space. This model is not a linear model. Thus, it can not be fitted by PROC REG or PROC GLM in SAS. The ML estimate of \( D \) is
\[ \hat{D} = \frac{\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j Y_{ij}}{\sum_i \sum_j \hat{\alpha}_i^2 \hat{\beta}_j^2} = \frac{\sum_i \sum_j (\bar{Y}_i - \bar{Y}_\cdot)(\bar{Y}_j - \bar{Y}_\cdot)Y_{ij}}{\sum_i \sum_j (\bar{Y}_i - \bar{Y}_\cdot)^2(\bar{Y}_j - \bar{Y}_\cdot)^2}. \]

Let
\[ SSAB^* = \sum_i \sum_j \hat{D}^2 \hat{\alpha}_i^2 \hat{\beta}_j^2. \]

Then,
\[ SSTO = SSA + SSB + SSAB^* + SSRem^* \]
and
\[ F^* = \frac{SSAB^*/1}{SSRem^*/(IJ - I - J)}, \]
follows \( F_{I,J-1,J-1} \) is \( H_0 : D = 0 \) is true. \( H_0 \) is rejected for large \( F^* \) value.
Multiple Comparison and Simultaneous Confidence Interval. Contrasts for Factor Effect. Simultaneous confidence intervals for

\[ D_{ii'}^\alpha = \alpha_i - \alpha_{i'}, \]
\[ D_{jj'}^\beta = \beta_j - \beta_{j'}, \]
can be computed, as well as for

\[ D_{ij,i'j'}^{(\alpha\beta)} = (\alpha\beta)_{ij} - (\alpha\beta)_{ij'} - (\alpha\beta)_{i'j} + (\alpha\beta)_{i'j'} \]

if there is replicated at least in one cell. For the balanced data, we assume that \( n_{ij} = K > 1 \) for all \( i \) and \( j \).

Tukey Procedure. Based on the theory introduced in one-way model, the simultaneous confidence interval for each \((i', i)\) is

\[ D_{ii'}^\alpha : \hat{D}_{ii'}^\alpha \pm T_\alpha \times std(\hat{D}_{ii'}^\alpha); \]
\[ D_{jj'}^\beta : \hat{D}_{jj'}^\beta \pm T_\beta \times std(\hat{D}_{jj'}^\beta); \]

and

\[ D_{ij,i'j'}^{(\alpha\beta)} : \hat{D}_{ij,i'j'}^{(\alpha\beta)} \pm T_{\alpha\beta} \times std(\hat{D}_{ij,i'j'}^{(\alpha\beta)}); \]

where

\[ T_\alpha = \frac{1}{\sqrt{2}} q_{I,IJ(K-1)}(1 - \alpha); \]
\[ T_\beta = \frac{1}{\sqrt{2}} q_{J,IJ(K-1)}(1 - \alpha); \]

and

\[ T_{\alpha\beta} = \frac{1}{\sqrt{2}} q_{(I-1)(J-1),IJ(K-1)}(1 - \alpha). \]

The null hypothesis, \( H_0 : \alpha_i = 0 \) for all \( i \); \( H_0 : \beta_j = 0 \) for all \( j \); or \( H_0 : (\alpha\beta)_{ij} = 0 \) for all \( i \) and \( j \) is rejected if one of the corresponding confidence intervals does not contain zero.

Scheffe Procedure. Based on the test for all contrasts equal zero, the Scheffe confidence interval is

\[ L_\alpha = \sum_i c_i^\alpha \alpha_i : \hat{L}_\alpha \pm S_\alpha \times std(\hat{L}_\alpha), \sum_i c_i^\alpha = 0; \]
\[ L_\beta = \sum_j c_j^\beta \beta_j : \hat{L}_\beta \pm S_\beta \times std(\hat{L}_\beta), \sum_j c_j^\beta = 0; \]

and

\[ L_{(\alpha\beta)} = \sum_i \sum_j c_{ij}^{(\alpha\beta)}(\alpha\beta)_{ij} : \hat{L}_{(\alpha\beta)} \pm S_{\alpha\beta} \times std(\hat{L}_{(\alpha\beta)}), \sum_i c_{ij}^{(\alpha\beta)} = \sum_j c_{ij}^{(\alpha\beta)} = 0, \]
where

\[ S^2_a = (I - 1)F_{I-1,IJ(n-1)}(1 - \alpha); \]

\[ S^2_\beta = (J - 1)F_{J-1,IJ(n-1)}(1 - \alpha); \]

and

\[ S^2_{a\beta} = (I - 1)(J - 1)F_{(I-1)(J-1),IJ(n-1)}(1 - \alpha). \]

**Bonferroni and Holm Procedure.** Bonferroni procedure tests all the selected (finite number) contrasts are zero. Holm Procedure tests those contrasts sequentially.

**Contrasts for Factor Cell Means.** If define

\[ D_{ij,i'j'} = \mu_{ij} - \mu_{i'j'}, \]

then, the Tukey simultaneous confidence interval is

\[ \hat{D}_{ij,i'j'} \pm T \times std(\hat{D}_{ij,i'j'}), \]

where

\[ T = \frac{1}{\sqrt{2}}q_{IJ,IJ(K-1)}(1 - \alpha). \]

The Scheffe’s procedure for

\[ L = \sum_i \sum_j c_{ij}\mu_{ij}, \sum_i \sum_j c_{ij} = 0, \]

is

\[ \hat{L} \pm S \times std(\hat{L}) \]

where

\[ S = (IJ - 1)F_{IJ-1,IJ(K-1)}(1 - \alpha). \]

**When One or Both Factors Quantitative.** When one is nominal variable and the other is continuous variable, the lack of fit model can be parallel here. If consider the interaction effect, the full model, is

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \]

and the reduced model is

\[ Y_{ijk} = \mu + \alpha_i + X_j\gamma_i + \epsilon_{ij}; \]

if do not consider interaction effect, the full model is

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \]
and the reduced model is

\[ Y_{ijk} = \mu + \alpha_i + X_j \gamma + \epsilon_{ij}. \]

There are several optional model when both factors are numeric variables, denote \( X_i^a \) and \( X_j^\beta \). One options is

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + X_i^a X_j^\beta \gamma + \epsilon \]

and \( \gamma \) is the index for the interaction effect, or

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + X_i^a X_j^\beta \gamma_1 + (X_i^a)^2 \gamma_2 + (X_j^\beta)^2 \gamma_3 + \epsilon. \]

These models are good options if there is no replication in each cell.

**Chapter 23: Two-Way ANOVA model (Unbalanced Case)**

When the data are unbalanced, the estimate in Table 8 is no longer correct. There is no explicit expression for the estimates. We have to use the coding method to fit the model by SAS. Look at the Growth Hormone Example on page 892, textbook. Let us denote \( \text{SSE}(X_1, \ldots, X_p) \) be the \( \text{SSE} \) of the model with \((X_1, \ldots, X_{p-1})\), where \( X_i \) can be either continuous variable or categorical variable.

**Type of Sum of Squares**

There are four types of sum of squares: Type I, Type II, Type III, and Type IV. We will learn three of them. Let us consider the two-way factor effect interaction model (model with factor \( A, B \), and \( A \times B \)),

\[ Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha \beta)_{ij} + \epsilon_{ijk}, \]

where \( k = 1, \ldots, n_{ij} \), \( \epsilon \sim N(0, \sigma^2) \), and

\[ \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha \beta)_{ij} = \sum_j (\alpha \beta)_{ij} = 0. \]

**Type I SS.** Assume the predictors are according to the order \( A, B, AB \).

\[ SS_1(AB) = SS(A, B) - SS(A, B, AB); \]

\[ SS_1(B) = SS(A) - SS(A, B); \]

and

\[ SS_1(A) = SS(\emptyset) - SS(A). \]

If the order is \( B, A, AB \), then

\[ SS_1(AB) = SS(A, B) - SS(A, B, AB); \]
\[ SS_1(A) = SS(B) - SS(A, B); \]

and

\[ SS_1(B) = SS(\phi) - SS(B). \]

Type II SS. The results of Type II SS does not depend on the order of the variable names. The definition is

\[ SS_2(AB) = SS(A, B) - SS(A, B, AB); \]
\[ SS_2(A) = SS(B) - SS(A, B); \]

and

\[ SS_2(B) = SS(A) - SS(A, B). \]

Type III SS. The results of Type III SS does not depend on the order of the variable names. If the interaction is in the model, then the definition is

\[ SS_3(AB) = SS(A, B) - SS(A, B, AB); \]
\[ SS_3(A) = SS(AB) - SS(A, B, AB); \]

and

\[ SS_3(B) = SS(AB) - SS(A, B, AB). \]

If the interaction is not in the model, then the definition is

\[ SS_3(A) = SS(B) - SS(A, B); \]

and

\[ SS_3(B) = SS(A) - SS(A, B). \]

The test for the significant based on the types of sums of squares is based on \( SSE \) in the full model.

Weighted and Unweighted Estimator for Treatment Effect. Assume that there are two treatments (\( A \) and \( B \)) and \( J \) blocks in the study. The objective is to find the treatment effect. Then, the model is

\[ Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \]

where \( \epsilon_{ijk} \sim N(0, \sigma^2) \), \( i = 1, 2(A, B), j = 1, \ldots, J, k = 1, \ldots, n_{ij} \) with \( n_{1j} = n_{2j} = n_j \). Let

\[ n = \sum_i \sum_j n_{ij} = 2 \sum_j n_j, \]

and \( \delta_j = \mu_{1j} - \mu_{2j} \). So \( \delta_j \) is the treatment effect of block \( j \). Then,

\[ \hat{\delta}_j = \bar{Y}_{1j} - \bar{Y}_{2j} \sim N(\delta_j, \frac{2\sigma^2}{n_j}). \]
If $\delta_j = \delta$ for all $j$, then a very good estimator for $\delta$, called weighted estimator, is

$$\hat{\delta}_W = \frac{\sum_j n_j \hat{\delta}_j}{\sum_j n_j} \sim N(\delta, \frac{4\sigma^2}{n}).$$

Another option estimator for $\delta$, called unweighted estimator, is

$$\hat{\delta}_U = \frac{1}{J} \sum \delta_j \sim N(\delta, \frac{2\sigma^2}{J^2} \sum_{j=1}^J \frac{1}{n_j}).$$

It is clear that when $n_1 = \cdots = n_J$, then

$$\frac{2}{J^2} \sum_{j=1}^J \frac{1}{n_j} = \frac{2}{Jn_1} = \frac{4}{n},$$

but if $n_i$ are not all equal, then

$$\frac{2}{J^2} \sum_{j=1}^J \frac{1}{n_j} > \frac{4}{n}.$$

Therefore, the test based on $\hat{\delta}_W$ is more powerful if there is no interaction effect.

If the interaction effect exists, then $\hat{\delta}_W$ is a biased estimator, but $\hat{\delta}_U$ is not, because based on factor effect model, the treatment effect is

$$\frac{1}{J} \sum \delta_j.$$

**Comment:** If type II sum of squares based on interaction model is significant, then the weighted test is significant.

**Multi-Way ANOVA model.** If there are more than two factors involved in the study, then a more complicated model should be established. Consider the three-way model,

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$

where $i = 1, \cdots, I$, $j = 1, \cdots, J$, $k = 1, \cdots, K$, $l = 1, \cdots, n_{ijkl}$, and $\epsilon_{ijkl} \sim N(0, \sigma^2)$ independently. The constrain for the parameter is still zero sum constrain. We call $(\alpha\beta\gamma)_{ijk}$ three-factor interaction effect; call $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$ tw-factor interaction effect; and call $\alpha_i$, $\beta_j$, $\gamma_k$ main effect.

There are several tasks of testing problem, such as test the interaction effects or main effects. We still can compute the simultaneous confidence intervals. I am going to concentrate a few examples in this section.