One Factor Experimental Studies

Suppose we have a finite population, with subjects labeled by \( s = 1, \ldots, N \). Assume that we can measure response on a subject without error, and represent the potential response of subject \( s \) under factor level \( j \) by \( y_{sj} \), where \( j = 1, \ldots, p \). Our interest is in comparing the average response in the population under different levels of the factor. Potentially, we could compute the mean response for the population under each factor level, i.e. \( \mu_j = \frac{1}{N} \sum_{s=1}^{N} y_{sj} \). Our interest is in comparison of these means.

The Experiment

Although all responses are potentially observable, we plan on observing response on different subsets of subjects under each of the different factor levels. Practically, we randomly permute the labels for the subjects, and then assign groups of subjects in the permutation to the factor levels. Let the first \( n_1 \) selected subjects constitute a sample for treatment 1, the next \( n_2 \) selected subjects constitute a sample for treatment 2, etc, until the last \( n_p = N - \sum_{j=1}^{p-1} n_j \) selected subjects constitute a sample for treatment \( p \). We assume that we use all subjects, so that \( N = \sum_{j=1}^{p} n_j \).

Random Variables

We represent the random variables for a random permutation of subjects by the vector \( \mathbf{Y} = \left( Y_{11}, Y_{21}, \ldots, Y_{N1} \right)' \) where \( Y_i \) represents response for the subject in position \( i \) in the permutation. Note that in this context, \( E(\mathbf{Y}) = \frac{1}{N} \mathbf{J}_N \mathbf{y} = \mathbf{1}_N \mu \) and \( \text{var}(\mathbf{Y}) = \sigma^2 \left( \mathbf{I}_N - \frac{\mathbf{J}_N}{N} \right) \). In order to capture the assignment of factor levels to positions, we represent the random variables using two subscripts, \( i \) and \( j \), where we index positions for factor level \( j \) by \( i = 1, \ldots, n_j \). Using this indexing, we represent the random variables in the form

\[
\left( Y_{11}, Y_{21}, \ldots, Y_{n_1}, Y_{12}, Y_{22}, \ldots, Y_{n_2}, \ldots, Y_{1p}, Y_{2p}, \ldots, Y_{n_p} \right)'.
\]
Models for Response

The statistical analysis is based on a comparison of various models for response. All models are of the form $Y_{ij} = E(Y_{ij}) + E_{ij}$. We consider three models:

1. $Y_{ij} = \mu_j + E_{ij}$ where $E(Y_{ij}) = \mu_j$
2. $Y_{ij} = \mu + E_{ij}$ where $E(Y_{ij}) = \mu$
3. $Y_{ij} = E_{ij}$ where $E(Y_{ij}) = 0$

The different models correspond to different assumptions about the expected value of $Y_{ij}$. We assume that the variance is the same for all models. This assumption implies that different levels of a factor only alter the expected value, not the variance.

For each model, we estimate the parameters in the model using ordinary least squares. Then, using the parameter estimates, we estimate the residuals. Our statistical comparison is based on a comparison of the sum of squared residuals. These sums are tabulated in an ANOVA table.

In order to interpret the comparison of the sum of squared residuals, we evaluate the expected value of the sum of squared residuals. We evaluate these expected values next. We simplify this evaluation by representing the models, estimators, and residuals using vectors.

Representation of Models using Vectors

Each model that we consider can be represented simultaneously for all random variables using vectors and matrices such that: $Y = X\beta + E$ where $E(Y) = X\beta$. Least squares estimators of $\beta$ are obtained by minimizing the sum of squared residual values,

$$E'E = (Y - X\beta)'(Y - X\beta) = \sum_{j=1}^{n_j} \sum_{i=1}^{n_i} E_{ij}^2.$$ This can be expanded such that

$$E'E = YY' - Y'X\beta - \beta'XY + \beta'XX\beta.$$ We differentiate this expression with respect to $\beta$, and obtain the least squares estimates of $\beta$ by setting the derivative to zero, resulting in the estimating equations given by $XX\hat{\beta} = XY$. Solving this expression (assuming $XX$ is non-singular) results in $\hat{\beta} = (XX)^{-1}XY$. We use the parameter estimates to form estimated residuals, $\hat{E} = Y - X\hat{\beta}$. The sums of squared values of these residuals are tabulated in an ANOVA table.

We can express the sum of squared residuals in a simpler form that makes evaluation of the expected value easier. Substituting the expression for the least squares estimator,

$$\hat{E} = Y - X\hat{\beta}$$

$$= Y - X(X'X)^{-1}XY$$

$$= (I_n - X(X'X)^{-1}X')Y$$
The matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ is called the “hat” matrix. We represent it by $\mathbf{H}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$. As a result, $\hat{\mathbf{E}} = (\mathbf{I}_N - \mathbf{H}_\mathbf{X}) \mathbf{Y}$. The hat matrix is symmetric (i.e. $\mathbf{H}_\mathbf{X}' = \mathbf{H}_\mathbf{X}$) and idempotent (i.e. $\mathbf{H}_\mathbf{X} \mathbf{H}_\mathbf{X} = \mathbf{H}_\mathbf{X}$). Also, $(\mathbf{I}_N - \mathbf{H}_\mathbf{X})$ is symmetric and idempotent.

As a result, we can express

$$Q = \hat{\mathbf{E}}' \hat{\mathbf{E}} = \mathbf{Y}'(\mathbf{I}_N - \mathbf{H}_\mathbf{X}) \mathbf{Y}.$$  

The sum of squared residuals is a scalar random variable which we represent by $Q$. Scalar random variables that can be represented by $\mathbf{Y}'\mathbf{A}\mathbf{Y}$ are called quadratic forms. Setting $\mathbf{A} = \mathbf{I}_N - \mathbf{H}_\mathbf{X}$, we see that $Q$ is a quadratic form. To evaluate the expected value of the sum of squared residuals, we evaluate $E(Q)$.

**Evaluating the Expected Value of a Quadratic Form**

We evaluate the expected value of the quadratic form $Q = \mathbf{Y}'\mathbf{A}\mathbf{Y}$, where $\mathbf{A}$ is symmetric and non-stochastic. For our evaluation, we assume that $E(\mathbf{Y}) = \mathbf{X}\beta$ and represent $\operatorname{var}(\mathbf{Y}) = \Sigma$. Since $Q$ is a scalar, $Q = \operatorname{Tr}(Q)$, and

$$Q = \operatorname{Tr}(\mathbf{Y}'\mathbf{A}\mathbf{Y}) = \operatorname{Tr}(\mathbf{A}\mathbf{Y}\mathbf{Y}').$$  

Now $\operatorname{var}(\mathbf{Y}) = E[(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)'] = \Sigma$ and hence $E[\mathbf{Y}\mathbf{Y}' - \mathbf{X}\beta\beta'\mathbf{X}'] = \Sigma$, or $E[\mathbf{Y}\mathbf{Y}'] = \Sigma + \mathbf{X}\beta\beta'\mathbf{X}'$ (as for example in Hocking, 1985, p27). Using this result,

$$E(Q) = \operatorname{Tr}[\mathbf{A}E(\mathbf{Y}\mathbf{Y}')] = \operatorname{Tr}[\mathbf{A}(\Sigma + \mathbf{X}\beta\beta'\mathbf{X}')].$$

$$= \operatorname{Tr}(\mathbf{A}\Sigma) + \beta'\mathbf{X}'\mathbf{A}\beta$$

Some other useful properties of quadratic forms are given by Hocking (1985, Corollary 2.2.1, p27).

**The ANOVA Table, Models, and Sums of Squares**

An ANOVA table summarizes the sum of squared residuals for different models. The differences between the models correspond to inclusion of additional parameters. Comparisons of these sums of squares are used to test the null hypotheses that the additional parameters are zero. We evaluate the expected value of the sum of squared residuals to interpret the comparisons.

For a 1 Factor study, the ANOVA table is given as follows. Assume $n$ subjects are randomly assigned to each of $p$ levels of a factor. Let $\mathbf{X} = \mathbf{I}_p \otimes \mathbf{I}_n$, $\mathbf{X}_1 = \mathbf{I}_p \otimes \mathbf{1}_n$, and $\mathbf{X}_2 = \left( \begin{array}{c} \mathbf{I}_{p-1}^T \mathbf{1}_n \end{array} \right) \otimes \mathbf{I}_n$. Also, note that $\mathbf{H}_\mathbf{X} = \mathbf{H}_{\mathbf{X}_1} + \mathbf{H}_{\mathbf{X}_2}$. Then
\[ Y'Y = Y'[I_{np} - H_{x_i}]Y + Y'H_{x_i}Y \]
\[ = \left( Y'[I_{np} - H_{x_1} - H_{x_2}]Y + Y'H_{x_1}Y \right) + Y'H_{x_i}Y \]

Table 1. Description of Sums of Squares in an ANOVA Table for a One Factor Study

<table>
<thead>
<tr>
<th>Description</th>
<th>Description</th>
<th>Sums of Squares (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Sums of Squares (corrected for mean)</td>
<td>Explained by the model ( Y_j = \mu_j + E_{ij} )</td>
<td>( Y'H_{x_i}Y )</td>
</tr>
<tr>
<td>Residual Sums of Squares</td>
<td>Residuals for the model ( Y_j = \mu_j + E_{ij} )</td>
<td>( Y'[I_{np} - H_{x}]Y )</td>
</tr>
<tr>
<td>Corrected Total Sums of Squares</td>
<td>Residuals for the model ( Y_j = \mu + E_{ij} )</td>
<td>( Y'[I_{np} - H_{x_i}]Y )</td>
</tr>
<tr>
<td>Mean Sum of Squares</td>
<td>Explained by the model ( Y_j = \mu + E_{ij} )</td>
<td>( Y'H_{x_i}Y )</td>
</tr>
<tr>
<td>Uncorrected Total Sums of Squares</td>
<td>Residuals for the model ( Y_j = E_{ij} )</td>
<td>( YY )</td>
</tr>
</tbody>
</table>

The models used in the ANOVA table can be represented in vector and matrix form. When sample sizes are not equal, the model \( Y_j = \mu_j + E_{ij} \) is given by

\[ Y = X\beta + E \]

where \( X = \bigoplus_{j=1}^{p} n_j \begin{pmatrix} 1_{n_j} & 0 & \cdots & 0 \\ 0 & 1_{n_j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{n_j} \end{pmatrix} \) and \( \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \).

For the model \( Y_j = \mu + E_{ij} \), the model is given by

\[ Y = X_1\mu + E \]

where \( X_1 = I_N \). For the model \( Y_j = E_{ij} \), there is no design matrix or parameters. These expressions can be used to evaluate the expected value of the sums of squares, and thus complete Table 2:
Table 2. Description of Sums of Squares in an ANOVA Table for a One Factor Study

<table>
<thead>
<tr>
<th>Description</th>
<th>Description</th>
<th>Sums of Squares ((Q))</th>
<th>Expected Value of the Sums of Squares ((E(\dot{Q})))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Sums of Squares</td>
<td>Explained by the model (Y_j = \mu_j + E_{ij})</td>
<td>(Y' H_{x_2} Y)</td>
<td></td>
</tr>
<tr>
<td>Residual Sums of Squares</td>
<td>Residuals for the model (Y_j = \mu_j + E_{ij})</td>
<td>(Y' [I_{np} - H_x] Y)</td>
<td></td>
</tr>
<tr>
<td>Corrected Total Sums of Squares</td>
<td>Residuals for the model (Y_j = \mu + E_{ij})</td>
<td>(Y' [I_{np} - H_x] Y)</td>
<td></td>
</tr>
<tr>
<td>Uncorrected Total Sums of Squares</td>
<td>Residuals for the model (Y_j = E_{ij})</td>
<td>(Y'Y)</td>
<td></td>
</tr>
</tbody>
</table>

**Simplifications when Responses Correspond to Ranks**

When the values for subjects are ranks (and there are no ties), \(\mu = \frac{N+1}{2}\) and \(\sigma^2 = \frac{N(N+1)}{12}\). Note that we will use simple (but less precise) notation of \(y_s\) as standing for the rank of the value of subject \(s\), instead of introducing new notation such as \(R(\dot{y}_s)\) for the rank of the value of subject \(s\).

**Analysis of Variance**

Suppose the treatments do not alter a subject’s response. The total (uncorrected sum of squares) is given by \(Y'Y = \sum_{i=1}^{N} Y_i^2\). In an ANOVA table, we partition the sums of squares.
Let $A_1 = \frac{J_N}{N} = N \frac{1_N}{N} \frac{1_N'}{N}$, $A_3 = X(X'X)^{-1}X' = \begin{pmatrix} \frac{J_{n_1}}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{J_{n_2}}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{J_{n_p}}{n_p} \end{pmatrix}$, and

$A_2 = A_3 - A_1 = \begin{pmatrix} \frac{J_{n_1}}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{J_{n_2}}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{J_{n_p}}{n_p} \end{pmatrix} - \frac{J_N}{N}$.

Then
### Description of Sums of Squares

**Due to the model**

\[ Y' A_2 Y = \sum_{j=1}^{p} n_j \left( \bar{Y}_j - \bar{Y} \right) = \sum_{j=1}^{p} \frac{1}{n_j} \left( \sum_{i=1}^{n_j} Y_{ij}^2 \right) - N\bar{Y} \]

**Residual**

\[ Y' [I_N - A_1] Y \]

**Corrected Total (for the mean)**

\[ Y' [I_N - A_1] Y = \sum_{i=1}^{N} \left( Y_{ij} - \bar{Y} \right)^2 = \sum_{i=1}^{N} \left( Y_i - \bar{Y} \right)^2 = (N-1)\sigma^2 \]

**Total Sums of Squares**

\[ Y'Y = (N-1)\sigma^2 + \mu^2 \]

Using the expressions corresponding to ranks,

**Description of Sums of Squares**

**Due to the model**

\[ Y' A_2 Y = \sum_{j=1}^{p} \frac{1}{n_j} \left( \sum_{i=1}^{n_j} Y_{ij}^2 \right) - \frac{N(N+1)^2}{4} \]

**Residual**

\[ Y' [I_N - A_1] Y \]

**Corrected Total (for the mean)**

\[ Y' [I_N - A_1] Y = \frac{N(N+1)(N-1)}{12} \]

**Total Sums of Squares**

\[ Y'Y = \frac{N(N+1)(2N-1)}{6} \]

Let us define the population variance as \[ \sigma^2 = \frac{N(N+1)}{12} \]. Since this is known, we can consider the statistic:

\[ H = \frac{Y' A_2 Y}{\sigma^2} = \frac{12}{N(N+1)} \sum_{j=1}^{p} \frac{1}{n_j} \left( \sum_{i=1}^{n_j} Y_{ij}^2 \right) - 3(N+1) \]

as following a chi-square distribution with \( p-1 \) degrees of freedom for large samples. This is the large sample Kruskal Wallis test statistic (see Hollander and Wolf, Nonparametric Statistical Methods, (1973) p115).

We can obtain the statistic by

1. Ranking the data overall
2. Run a 1-way ANOVA table on the ranked data
3. Evaluate \( \sigma^2 \) by dividing the corrected total sums of squares by \( N-1 \)
4. Dividing the model SS by \( \sigma^2 \), and comparing it with a Chi-Square distribution with \( p-1 \) degrees of freedom.