Analysis of a One Factor Experiment

Ed Stanek

Introduction

We derive the least squares estimators and the ANOVA table for a one factor experimental design with equal numbers of units, \( i = 1, \ldots, n \) assigned to each level (or treatment), \( j = 1, \ldots, p \) of the factor. We begin with a model for random variables for the design, express this model using matrix algebra, and minimize the sum of squared error, and summarize the results in an ANOVA table. These results are given in many places (see for example, Kirk (1995), p224-226, Searle, (1987), p253-265.

Study Population and Parameters

We assume the study is conducted using a large population of \( s = 1, \ldots, N \) subjects. Potentially, each subject could receive any level of the factor. We represent the potentially observable response of a subject after treatment as a fixed constant \( \mu_j \). Using this representation of response, we define parameters for the mean and variance of treatment \( j \) as

\[
\mu_j = \frac{1}{N} \sum_{s=1}^{N} \mu_s \\
\sigma_j^2 = \left( \frac{N-1}{N} \right) \left( \frac{N-1}{N} \right) \left( \frac{\sum_{s=1}^{N} (\mu_s - \mu_j)^2}{N-1} \right)
\]

for \( j = 1, \ldots, p \). We assume that the population size is very large so that we can approximate \( \frac{N-1}{N} \approx 1 \).

The Experiment

The experiment is conducted by randomly assigning \( i = 1, \ldots, n \) subjects to each treatment, administering the treatment, and then observing response. The index \( i \) represents the position of the subject in the sample for each treatment. Since the subject assigned to position \( i \) for treatment \( j \), is random, we represent response as the random variable

\[ Y_{ij} = \text{response for the subject assigned to position } i \text{ for treatment } j. \]

The response can also be represented via the model:

\[ Y_{ij} = \mu_j + E_{ij} \]

where \( E_{ij} \) is a random variable that represents the difference between response for the subject assigned to position \( i \) for treatment \( j \) and the mean response for treatment \( j \). Since all subjects
have an equal chance of being assigned to treatment \( j \), \( E(Y_{ij}) = \mu_j \) and \( E(E_{ij}) = 0 \), where \( E \) denotes the expected value over all possible random assignments. Also, \( \text{var}(E_{ij}) = \sigma_j^2 \). We assume that \( N \) is much larger than \( n \) so that \( \text{cov}(E_{ij}, E_{ir}) = 0 \) for \( i \neq i^* \), and that \( \sigma_j^2 = \sigma^2 \) for all \( j = 1, \ldots, p \).

**Vector and Matrix Representation of the Model**

We represent response for the \( n \) subjects assigned treatment \( j \) by the vector

\[
Y_j' = \left( Y_{1j} \ Y_{2j} \ \cdots \ Y_{nj} \right),
\]

and the response vector for the study as \( Y_{np \times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} \), and define similar vectors for \( E_j \) and \( E \). Using this notation, we can summarize the model as

\[
Y = X\beta + E
\]

where \( X_{np \times p} = I_p \otimes I_n \) and \( \beta' = (\mu_1 \ \mu_2 \ \cdots \ \mu_p) \). Also, note that \( \text{var}(Y) = \text{var}(E) = \sigma^2 I_{np} \).

**Obtaining Estimates of \( \beta \) that minimize the sum of the squared Error.**

We determine estimates of \( \beta \) that minimize the sum of squared error. Since the error is defined as \( E = Y - X\beta \), the sum of squared errors is given by

\[
\sum_{i=1}^{n} \sum_{j=1}^{p} (Y_{ij} - \mu_j)^2 = E'E = (Y - X\beta)'(Y - X\beta) = (Y' - \beta'X')(Y - X\beta) = Y'Y - \beta'X'Y + \beta'X'\beta = Y'Y - 2Y'X\beta + \beta'X'\beta
\]

Each term in this expression is a scalar.

Our objective is to estimate \( \beta \) such that this function is minimized. To do so, we differentiate with respect to \( \beta \), and then set the resulting derivative to zero. Since the function is quadratic in \( \beta \), when the derivative is zero, we will have a relative minimum (since the second derivative is positive). Using basic differentiation,

\[
\frac{\partial (E'E)}{\partial \beta} = 2X'Y + 2X'X\beta.
\]

Setting the derivative to zero, \( X'X\hat{\beta} = X'Y \), and since \( X'X \) is non-singular,
\( \hat{\beta} = (XX)^{-1} XY. \)

We can determine an explicit expression for the least squares estimates by noting that

\[
(X'X)^{-1} = \frac{1}{n} I_p \quad \text{and} \quad X'Y = \begin{pmatrix}
\sum_{i=1}^{n} Y_{i1} \\
\vdots \\
\sum_{i=1}^{n} Y_{ip}
\end{pmatrix}
\]

so that \( \hat{\beta} = \left( (\bar{y}) \right) \). Finally,

\[
\text{var}(\hat{\beta}) = \text{var}\left( \left( (XX)^{-1} X' \right) Y \right)
\]

\[
= \left( (XX)^{-1} X' \right) \text{var}(Y) \left( (XX)^{-1} X' \right)'
\]

\[
= \left( (XX)^{-1} X' \right) I_p \sigma^2 X (XX)^{-1}
\]

\[
= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1}
\]

\[
= \sigma^2 (X'X)^{-1}
\]

Parameterizations

Parameters in the model may be represented in different ways. For example, defining \( \mu = \frac{1}{3} \sum_{j=1}^{3} \mu_j \), we can represent the parameters \( \beta' = (\mu_1, \mu_2, \mu_3) \) by the equivalent set of parameters given by \( \tau' = (\mu, \tau_1, \tau_2) \), where \( \tau_1 = \mu_1 - \mu \) and \( \tau_2 = \mu_2 - \mu \). Generally, such re-parameterizations can be represented as a non-singular transformation of the parameters where \( P\beta = \tau \), and \( P \) is a non-singular transformation matrix. In the example given above,

\[
P = \frac{1}{3} \begin{pmatrix}
1 & 1 & 1 \\
2 & -1 & -1 \\
-1 & 2 & -1
\end{pmatrix}
\]

Sums of Squares and the ANOVA Table

The results of fitting a model are commonly summarized in an analysis of variance table. The objective of the table is to present the extent to which the sum of the squared residuals are reduced by addition of terms in the model. It is also possible to determine the expected value of the sums of squares, and hence estimate the contribution to the reduction due to additional parameters. With additional distributional assumptions, a null distribution can be assumed and used for traditional hypothesis tests.
Residuals are expressed as the difference between the observed data and the predicted values, \( Y - X\hat{\beta} \). Since \( \hat{\beta} = (XX)^{-1}XY \), this expression is equal to

\[
Y - X(XX)^{-1}XY = \left[ I_{np} - X(XX)^{-1}X' \right]Y.
\]

The residuals are unchanged by a re-parameterization of the model. One such re-parameterization will result in deviations from means parameters, with the design matrix given by

\[
X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \text{ where } X_1 = 1_{np}, \quad \text{and} \quad X_2 = \begin{pmatrix} I_{p-1} \\ -1_{p-1} \end{pmatrix} \otimes 1_n.
\]

Note that with this parameterization, \( XX = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix} \). As a result, the residuals can be expressed as

\[
I_{np} - X_1 \left( X_1'X_1 \right)^{-1}X_1' - X_2 \left( X_2'X_2 \right)^{-1}X_2' Y.
\]

Using this expression for the residuals, and noting that \( X_1'X_2 = 0 \), the sum of squared residuals simplifies to

\[
SSR = Y' \left[ I_{np} - X_1 \left( X_1'X_1 \right)^{-1}X_1' \right] Y - Y'X_2 \left( X_2'X_2 \right)^{-1}X_2' Y.
\]

The residuals can be summarized in a table as follows:

<table>
<thead>
<tr>
<th>Description</th>
<th>Sums of Squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to the model</td>
<td>( Y'X_2 \left( X_2'X_2 \right)^{-1}X_2'Y )</td>
</tr>
<tr>
<td>Residual</td>
<td>( SSR )</td>
</tr>
<tr>
<td>Corrected Total (for the mean)</td>
<td>( Y' \left[ I_{np} - X_1 \left( X_1'X_1 \right)^{-1}X_1' \right] Y )</td>
</tr>
<tr>
<td>Total Sums of Squares</td>
<td>( YY )</td>
</tr>
</tbody>
</table>
Reference