20 Analysis of Variance

Suppose we have a field trial in which various types of treatments have been tried on different subjects and the effects recorded as an observation x for each individual. There are n_i individuals with treatment i and the observations from them are $x_{i,1}, \ldots, x_{i,n_i}$. We have k such sets of observatons of sizes $n_1, \ldots, n_k$ respectively, for a total of $N = n_1 + \cdots + n_k$ observations. The model assumes that each x_{i,n_i} is normally distributed with mean μ_i due to the effect of the i-th treatment and they all have a common variance σ^2. The null hypothesis is that there is no difference between the treatments or equivalently $\mu_1 = \mu_2 = \cdots = \mu_k$. The loglikelihood under the null hypothesis is

$$\log L_0 = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\mu)^2$$

Maximization with respect to σ and μ yields

$$\begin{aligned}
\hat{\mu} &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{i,j} \\
\hat{\sigma}_0^2 &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\hat{\mu})^2 \\
\log \hat{L}_0 &= -\frac{N}{2}\log 2\pi - N\log\hat{\sigma}_0 - \frac{N}{2}
\end{aligned}$$

Similarly under H_1,

$$\begin{aligned}
\log L_1 &= -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\mu_i)^2 \\
\hat{\mu}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j} \\
\hat{\sigma}_1^2 &= \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\hat{\mu}_i)^2 \\
\log \hat{L}_1 &= -\frac{N}{2}\log 2\pi - N\log\hat{\sigma}_1 - \frac{N}{2}
\end{aligned}$$

The loglikelihood ratio criterion takes the form

$$-2\log\frac{L_0}{L_1} = N\log\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}$$

so that a test can be based on

$$U = \frac{\hat{\sigma}_0^2 - \hat{\sigma}_1^2}{\hat{\sigma}_1^2}$$

A computation yields

$$\begin{aligned}
N\hat{\sigma}_0^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\hat{\mu})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{i,j}^2 - 2\hat{\mu}\sum_{j=1}^{n_i}x_{i,j} + N\hat{\mu}^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{i,j}^2 - N\hat{\mu}^2
\end{aligned}$$

On the other hand

$$\begin{aligned}
N\hat{\sigma}_0^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\hat{\mu})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\hat{\mu}_i)^2 + 2\sum_{i=1}^{k}(\hat{\mu}_i-\hat{\mu})\sum_{j=1}^{n_i}(x_{i,j}-\hat{\mu}) + \sum_{i=1}^{k}n_i(\hat{\mu}_i-\hat{\mu})^2 \\
&= N\hat{\sigma}_1^2 + \sum_{i=1}^{k}n_i(\hat{\mu}_i-\hat{\mu})^2 \\
&= N\hat{\sigma}_1^2 + \sum_{i=1}^{k}n_i\hat{\mu}_i^2 - N\hat{\mu}^2
\end{aligned}$$

The following quantities are to be computed.

$$T_i = \sum_{j=1}^{n_i} x_{i,j}$$

$$T = \sum_{i=1}^{k} T_i$$

$$c = \frac{T^2}{N}$$

$$A = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j}^2$$

$$B = \sum_{i=1}^{k} \frac{T_i^2}{n_i}$$

Then

$$A = c + (B - c) + (A - B)$$

$$N\hat{\sigma}_0^2 = A - c$$

and

$$N\hat{\sigma}_1^2 = (A - B)$$

so that

$$U = \frac{B - c}{A - B}$$

It can be seen that under the null hypothesis $(B - c)$ and $(A - B)$ are independent $\sigma^2\chi^2$ with $(k - 1)$ and $(N - k)$ degrees of freedom and

$$F = U\frac{N - k}{k - 1} = \frac{\frac{B-c}{k-1}}{\frac{A-B}{N-k}}$$

is an $F_{k-1,N-k}$. A large values of F leads to the rejection of H_0 that $\mu_1 = \cdots = \mu_k$.

Some times the population on which the treatments are tried is not uniform. For example each treatment i could be tried on the j-th group once leading to an observation $x_{i,j}$. The number of treatments is k and the number of groups is n, each group consisting of k observations. The total number of observations is $N = kn$. The model is that the observation $x_{i,j}$ is normally distributed with mean $\mu_i + a_j$ and variance σ^2. Actually there is a redundancy of parameters here, and it is better to write the mean as $\mu + \mu_i + a_j$ with $\sum_i \mu_i = \sum_j a_j = 0$ with a total of $n + k - 1$ parameters. We are interested in testing the null hypothesis $\mu_1 = \cdots = \mu_k = 0$ and we really do not care about $a_1, \ldots, a_n$. With similar calculations we obtain

$$\hat{\mu}_i = \frac{1}{N} \sum_{i,j} x_{i,j}$$

$$\hat{\mu}_i = \frac{1}{n} \sum_{j} x_{i,j} - \hat{\mu}$$

$$\hat{a}_j = \frac{1}{k} \sum_{i} x_{i,j} - \hat{\mu}$$

$$\sum_{i,j} x_{i,j}^2 = N\sigma_1^2 + (k \sum_i \hat{\mu}_i^2 - N\hat{\mu}^2) + (n \sum_j \hat{a}_j^2 - N\hat{\mu}^2) + N\hat{\mu}^2$$

If we define as before

$$T_i = \sum_{j=1}^{n} x_{i,j}$$

$$S_j = \sum_{i=1}^{k} x_{i,j}$$

$$T = \sum_{i=1}^{k} T_i = \sum_{j=1}^{n} S_j$$

$$c = \frac{T^2}{N}$$

$$A = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{i,j}^2$$

$$B = \frac{1}{n} \sum_{i=1}^{k} T_i^2$$

$$C = \frac{1}{k} \sum_{j=1}^{n} S_j^2$$

We see that

$$A = c + (B - c) + (C - c) + E = Q_1 + Q_2 + Q_3 + Q_4$$

where $E = N\sigma_1^2$ and $N\sigma_0^2 - N\sigma_1^2 = B - c$. The ratio

$$F = \frac{\frac{B-c}{k-1}}{\frac{E}{(n-1)(k-1)}}$$

is an $F_{k-1,(n-1)(k-1)}$. The proof that the various components of the sum of squares are independent χ^2s depends on two observations. First each term is of the form

$$Q_r = \|\tilde{x}\|^2 - \inf_{y \in X_r} \|\tilde{x} - y\|^2$$

where $\tilde{x}$ is the $N = nk$ dimensional vector $\{x_{i,j}\}$ and $\{X_r : r = 1, 2, 3, 4\}$ are orthogonal subspaces of R^N. If we show that X_1, X_2, X_3 are mutually orthogonal then clearly X_4 is the orthogonal complement of $X_1 \oplus X_2 \oplus X_3$. It is easily verified that

$$X_1 = \{\tilde{x} : x_{i,j} \equiv a \text{ for all } i, j\}$$

$$X_2 = \{\tilde{x} : x_{i,j} \equiv a_i \text{ for all } i, j \text{ with } \sum_{i=1}^{k} a_i = 0\}$$

$$X_3 = \{\tilde{x} : x_{i,j} \equiv a_j \text{ for all } i, j \text{ with } \sum_{j=1}^{n} a_j = 0\}$$

and that they are mutually orthogonal.

21 General Linear Models

General linear model is of the following form. There are unknown parameters $\sigma^2, \theta_1, \ldots, \theta_k$ and observations $x_1, x_2, \ldots, x_n$. The x_i are assumed to be independent and normally distributed with mean

$$E[x_i] = \sum_{j=1}^{k} a_{i,j}\theta_j$$

and variance σ^2. The factors $\{a_{i,j}\}$ are assumed to be known constants. The matrix A is assumed to be of rank k (otherwise we can reduce the number of real parameters). The null hypothesis is that $\theta \in \Theta_0$, a linear subspace (hyperplane) of R^k of dmension $r < k$, specified by $k - r$ linear relations. The analysis depends on the two quantities

$$Q_1 = \inf_{\theta \in R^k} \|x - A\theta\|^2$$
$$Q_0 = \inf_{\theta \in \Theta_0} \|x - A\theta\|^2$$

The ratio

$$F = \frac{\frac{Q_0 - Q_1}{k-r}}{\frac{Q_1}{n-k}}$$

is an $F_{k-r,n-k}$. The actual minimization involves inverting the matrix A^*A for the computation of Q_1 and a similar one for the computation of Q_0. In the examples we discussed this is particularly easy.

Rather than discuss the general theory we will do some examples.