Spring 2015 Statistics 151a (Linear Models) : Lecture Ten

Aditya Guntuboyina

19 February 2015

1 One Way Analysis of Variance

Consider the model

$$y_{ij} = \mu_i + e_{ij} \qquad \text{for } i = 1, \ldots, t \text{ and } j = 1, \ldots, n_i$$

where e_{ij} are i.i.d normal random variables with mean zero and variance σ^2. Let $\sum_{i=1}^{t} n_i = n$.

This model is used for the following kinds of situations:

1. There are t treatments and n subjects. Each subject is given one (and only one) of the j treatments. $y_{i1}, \ldots, y_{in_i}$ denote the scores of the subjects that received the ith treatment.
2. We are looking at some performance of n subjects who can naturally be divided into t groups. We would like to see if the performance difference between the subjects can be explained by the fact that there in these different groups. $y_{i1}, \ldots, y_{in_i}$ denote the performance of the subjects in the ith group.

Often this model is also written as

$$y_{ij} = \mu + \tau_i + e_{ij} \qquad \text{for } i = 1, \ldots, t \text{ and } j = 1, \ldots, n_i \tag{1}$$

where μ is called the baseline score and τ_i is the difference between the average score for the ith treatment and the baseline score. In this model, μ and the individual τ_is are not estimable. It is easy to show that here a parameter $\lambda\mu + \sum_{i=1}^{t} \lambda_i \tau_i$ is estimable if and only if $\lambda = \sum_{i=1}^{t} \lambda_i$. Because of this lack of estimability, people often impose the condition $\sum_{i=1}^{t} \tau_i = 0$. This condition ensures that all parameters μ and $\tau_1, \ldots, \tau_t$ are estimable. Moreover, it provides a nice interpretation. μ denotes the baseline response value and τ_i is the value by which the response value needs to be adjusted from the baseline μ for the group i. Because $\sum_i \tau_i = 0$, some adjustments will be positive and some negative but the overall adjustment averaged across all groups is zero.

How does one test the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ in this model? This is simply a linear model and we can therefore use the F-test. We just need to find the RSS in the full model (M) and the RSS in the reduced model (m). What is the RSS in the full model? Let $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $\bar{y} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}/n$. Write

$$\begin{aligned}
\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \mu_i)^2 \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + 2 \sum_{i=1}^{t} (\bar{y}_i - \mu_i) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) + \sum_{i=1}^{t} n_i (\bar{y}_i - \mu_i)^2 \\
&= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{t} n_i (\bar{y}_i - \mu_i)^2 .
\end{aligned}$$

Therefore, the least squares estimate of μ_i is $\hat{\mu}_i = \bar{y}_i$. If we write μ_i as $\mu + \tau_i$ with $\sum_i \tau_i = 0$, then the least squares estimate of μ is $\bar{y}$ and the least squares estimate of τ_i is $\bar{y}_i - \bar{y}$.

The RSS in the full model is

$$RSS(M) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 .$$

Check that the RSS in the reduced model is

$$RSS(m) = \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{t} n_i (\bar{y}_i - \bar{y})^2 .$$

Thus the F-statistic for testing $H_0 : \mu_1 = \cdots = \mu_t$ is

$$T = \frac{\sum_{i=1}^{t} n_i (\bar{y}_i - \bar{y})^2/(t-1)}{\sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2/(n-t)}$$

which has the F-distribution with $t-1$ and $n-t$ degrees of freedom under H_0.

2 Two-way Analysis of Variance with No Replication

Consider the same setting as last section where there are t treatments and we are interested in checking if they have equal effects or if there are any differences. However we cannot usually ignore other major factors that may contribute significantly to the total variability. For this, one often separates the experimental units into groups, called blocks which are homogeneous with respect to all non-treatment factors and replicate the complete set of t treatments inside each block.

Simplest model for this is the following where we have $n = tb$ experimental units and these are divided into b blocks of t units each. We then assign the t treatments one to each unit inside each block so that each block represents a complete replication of the treatments. Let y_{ij} denote the yield on the ith treatment inside the jth block with $i = 1, \ldots, t$ and $j = 1, \ldots, b$. A simple additive effects model here is:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \qquad \text{for } i = 1, \ldots, t \text{ and } j = 1, \ldots, b$$

where $\{\epsilon_{ij}\}$ denote i.i.d normal random variables with mean zero and variance σ^2. We make the assumption $\sum_i \tau_i = 0$ and $\sum_j \beta_j = 0$ in order to ensure that all parameters $\mu, \tau_1, \ldots, \tau_t, \beta_1, \ldots, \beta_b$ are estimable. With these assumptions, μ can be interpreted as the overall mean effect, τ_i is the effect of the ith treatment and β_j is the effect of the jth block.

The hypothesis of no difference in the treatments is $H_0 : \tau_1 = \cdots = \tau_t = 0$. How to test this?

Because of the assumptions $\sum_i \tau_i = 0$ and $\sum_j \beta_j = 0$, observe that we can decompose

$$\sum_{i=1}^{t} \sum_{j=1}^{b} (y_{ij} - \mu - \tau_i - \beta_j)^2$$

as

$$\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + b \sum_{i=1}^{t} (\bar{y}_{i.} - \bar{y}_{..} - \tau_i)^2 + t \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..} - \beta_j)^2 + tb(\bar{y}_{..} - \mu)^2.$$

Here $\bar{y}_{i.} = \sum_{j=1}^{b} y_{ij}/b$, $\bar{y}_{.j} = \sum_{i=1}^{t} y_{ij}/t$ and $\bar{y}_{..} = \sum_i \sum_j y_{ij}/(tb)$.

It immediately follows therefore that

$$RSS(M) = \sum_{i=1}^{t} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

and that

$$RSS(m) = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + b\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2 .$$

The residual degrees of freedom in M is $n-p-1$ with $n = tb$ and $p = t+b-1$. This equals $(t-1)(b-1)$. The quantity q in the F-test equals b. Thus the p-value for the F-test is given by

$$\mathbb{P}\left\{ F_{t-1,(t-1)(b-1)} > \frac{b\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})}{\sum_i \sum_j (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2} \frac{(t-1)(b-1)}{(t-1)} \right\}.$$