Fall 2013 Statistics 151 (Linear Models) : Lecture Eleven

Aditya Guntuboyina

03 October 2013

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 Permutation Tests

We have studied hypothesis testing in the linear model via the F-test so far. Suppose we want to test a linear hypothesis about $\beta = (\beta_0, \ldots, \beta_p)^T$ in the full linear model (denoted by M). We first construct a reduced model which incorporates the hypothesis in the full model M. Call this reduced model m. We then look at the quantity:

$$T := \frac{(RSS(m) - RSS(M))/(p-q)}{RSS(M)/(n-p-1)}.$$

It makes sense to reject the null hypothesis if T is large. To answer the question: *how large is large?*, we rely on the assumption of normality of the errors i.e., $e \sim N(0, \sigma^2 I)$ to assert that $T \sim F_{p-q,n-p-1}$ under H_0. As a result, a p-value can be obtained as $\mathbb{P}\{F_{p-q,n-p-1} > T\}$.

Suppose we do not want to assume normality of errors. Is there any way to obtain a p-value? This is possible in some cases via permutation tests. We provide two examples below.

2.1 Testing for all explanatory variables

We want to test the null hypothesis that all explanatory variables can be thrown away without assuming that $e \sim N(0, \sigma^2 I)$. Under the null hypothesis, we assume that if the response variable y has no relation to the explanatory variables. Therefore, it is plausible to assume that under the null hypothesis, the values of the response variable $y_1, \ldots, y_n$ are randomly distributed between the n subjects without relation to the predictors. This motivates the following test:

1. Randomly permute the response values: $y_1, \ldots, y_n$.
2. Calculate the quantity
$$\frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}.$$
with the response values being the permuted values in the pervious step.
3. Repeat the above pair of steps a large number of times.

4. This results in a large number of values of the test statistic (one for each permutation of the response values). Let us call them $T_1, \ldots, T_N$. The p-value is calculated as the proportion of $T_1, \ldots, T_N$ that exceed the original test statistic value T (T is calculated with the actual unpermuted response values $y_1, \ldots, y_n$).

The idea behind this test is as follows: From the given data, we calcuate the value of

$$\frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}.$$

We need to know how extreme this value is under the null hypothesis. Under the assumption of normality, we can assess this by the F-distribution. But we need to do this without assuming normality. For this, we try to generate values of this quantity under the null hypothesis. The idea is to do this by calculating the statistic after permuting the response values. Because once the response values are permuted, all association between the response and explanatory variables breaks down so that the values of

$$\frac{(RSS(m) - RSS(M))/p}{RSS(M)/(n-p-1)}.$$

for the permuted response values resembles values generated under the null hypothesis. The p-value is then calculated as the proportion of these values larger than the observed value.

2.2 Testing for a single explanatory variable

How do we test if, say, the first explanatory variable is useful? We calcuate the t-statistic:

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}$$

and calculate p-value by comparing it with the t_{n-p-1} distribution (which requires normality). How to do this without normality?

We can follow the permutation test by permuting the values of x_1. For each permutation, we calculate the t-statistic and the p-value is the proportion of these t-values which are larger than the observed t-value in absolute value.