Spring 2015 Statistics 151 (Linear Models) : Lecture Eleven

Aditya Guntuboyina

24 February 2015

1 Two-way Analysis of Variance with No Replication

Consider the same one way analysis of variance setting 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

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

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.