Nested Models

May 12, 2020

1 L10: Nested Models

1.1 2 Examples in the lecture notes

1.1.1 Example 1) Quality control raw materials

```
In [1]: purity <- c(1,-1,0,-2,-3,-4,-2,0,1,1,4,0,
                    1,-2,-3,0,4,2,-1,0,-2,0,3,2,
                    2,4,0,-2,0,2,1,-1,2,3,2,1)

In [2]: supplier <- factor(c(rep(1:3, rep(12,3))))

In [3]: batch <- factor(c(rep(rep(1:4, rep(3,4)),3)))

In [4]: qc <- data.frame(purity, supplier, batch)

In [5]: rm(purity, supplier, batch)
        attach(qc)
```

1.2 Remember, the ANOVA R output for a nested design is calibrated for A & B both being fixed.

If:

1. A is fixed and B is random;
2. A and B are both random,

then you have to generate your F statistic slightly differently.

In this case, A represents the supplier. **A is fixed.** In our example the number of factor A treatments, $a = 3$ (three suppliers).

B represents the batch - let's treat this as random. In our example, the number of factor B treatment levels within each Factor A, $b = 4$ (four batches per supplier).

1.2.1 Hypotheses and assumptions

Since A is treated as being fixed and B as random, here are our assumptions:

$\sum_{i=1}^{a} \tau_i = 0$, for $i = 1, 2, 3$

$\beta_{(i)j} \sim \text{NID}(0, \sigma_B^2)$

Therefore our two hypothesis, that there are no differences between the treatments means, and therefore that the treatment effects are zero:

$H_{0A} : \tau_1 = \tau_2 = \tau_3 = 0,$

where the alternative hypothesis H_{1A} is that at least one of the treatment effects is non-zero.

$H_{0B} : \sigma_B^2 = 0,$

where the alternative hypothesis H_{1B} is that the variance for the $\beta_{(i)j}$ random variable is non-zero, i.e. that there is indeed some variation in between Factor B levels within each Factor A level.

Because of the way the expectations for the mean squares works, the F-test for factor A is:

$$F_A = \frac{MS_A}{MS_{(A)B}},$$

1.3 Degrees of Freedom:

- A has $(a - 1) = 3 - 1 = 2$ degrees of freedom
- (A)B has $a(b - 1) = 3(4 - 1) = 9$ degrees of freedom
- Therefore our hypothesis test for A, where the null hypothesis is that there are no differences in treatment effects between Factor A levels (i.e. between suppliers), can be rejected at the $100\alpha\%$ significance level if:

  $F_A > F_{2,9,\alpha}$

NOT $F_A = \frac{MS_A}{MS_R}$ **as is usually calculated in a two-way ANOVA in R.**

---

2 IMPORTANT

NESTED FACTORS ARE ENTERED INTO R VIA `A/B` SYNTAX

```
In [13]: ## AHHH NICE TRICK WITH R ANOVA
         qc.aov <- aov(purity ~ supplier/batch)

         summary(qc.aov)
```

```
               Df Sum Sq Mean Sq F value Pr(>F)
supplier        2  15.06   7.528   2.853 0.0774 .
supplier:batch  9  69.92   7.769   2.944 0.0167 *
Residuals      24  63.33   2.639
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
In [32]: MS.A <- 7.528
         MS.A.B <- 7.769
         F.A <- MS.A / MS.A.B
```

```
In [15]: F.A
```

0.968979276612177

```
In [16]: pf(F.A,2,9)
```

0.584206149199484

Therefore the p-value of the supplier differences hypothesis test is actually far less significant than the default R input. What looked like it was nearly in the 5% significance level is actually no where near.

We do not have sufficient evidence to reject the null hypothesis, and therefore accept that there are no significant differences between suppliers.

Looking at the p-value associated with F_B however shows there to be significant variation between batch levels within a supplier group. We reject H_{0B} at the 5% level of significance, accepting the alternative hypothesis that there's at least one batch within a supplier group that has a non-zero treatment effect.

We can estimate the variance of the $\beta_{(i)j}$ random variable via:

$$\hat{\sigma}_B^2 = \frac{MS_{(A)B} - MS_R}{n}$$

```
In [18]: var.B <- (7.769 - 2.639) / 3
```

```
         var.B
```

1.71

```
In [27]: model.tables(qc.aov, type='means')
```

```
Tables of means
Grand mean

0.3611111

 supplier
supplier
      1       2       3
-0.4167  0.3333  1.1667
```

```
 supplier:batch
        batch
supplier 1       2       3       4
       1  0.0000 -3.0000 -0.3333  1.6667
       2 -1.3333  2.0000 -1.0000  1.6667
       3  2.0000  0.0000  0.6667  2.0000
```

2.0.1 Parameter estimates:

Recall:

$$\hat{\tau}_i = \bar{y}_{i..} - \bar{y}_{i..}$$

```
In [30]: y.bar.1.. <- -0.4167
         y.bar.2.. <- 0.3333
         y.bar.3.. <- 1.1667

         y.bar... <- 0.3611111
```

```
In [31]: tau.hat <- c(y.bar.1.., y.bar.2.., y.bar.3..) - y.bar...
         tau.hat
```

1. -0.7778111 2. -0.0278111 3. 0.8055889

```
In [10]: model.tables(qc.aov, type='effects')
```

```
Tables of effects

 supplier
supplier
      1       2       3
-0.7778 -0.0278  0.8056

 batch
batch
      1       2       3       4
-0.1389 -0.6944 -0.5833  1.4167
```

2.1 Example 2: Effect of daylight on growth of plants

Experiment to measure the effect that exposure to daylight has on the growth of plants

- 6 different conditions corresponding to 3 different exposures of sunlight (8, 12, 16 hours), at two glasshouse locations (low and high nighttime temperatures) -> these is a fixed factor.
- 3 randomly assigned pots per condition -> this is a random condition.
- 4 plants per pot per condition.

  CG NOTE: if you hear language like treatment B **PER** treatment A, then you should consider this nesting

  Also look out in these scenarios for which are the fixed and which are the random experiemental factors

2.1.1 Factors

- **A**: 6 fixed conditions, $\sum_{i=1}^{6} \tau_i = 0$, $i = 1, 2, 3, 4, 5, 6$. $a = 6$
- **B**: 3 random pots within each condition, $\beta_{(i)j} \sim \text{NID}(0, \sigma_B^2)$. $b = 3$

2.1.2 Characterisation of the experimental design

- We have a nested design;
- There are $abn = 6 \times 3 \times 4 = 72$ observations in total
- There should be equal numbers of n replicates in each jth level of factor B, within each ith level of factor A.

```
In [53]: growth <- c(35, 40, 30, 45, 25, 45, 55, 50, 30,
                     30, 25, 30, 50, 55, 40, 35, 35, 35, 30,
                     40, 45, 40, 40, 50, 50, 45, 50, 45, 55,
                     60, 50, 50, 55, 45, 65, 55, 85, 60, 90,
                     85, 65, 70, 80, 65, 70, 70, 70, 70, 60,
                     55, 35, 70, 60, 85, 45, 75, 65, 65, 85, 75,
                     70, 90, 85, 85, 60, 70, 70, 70, 110, 70, 90, 80)/10
```

2.2 Coding up the nested factors

- We have 72 data points where the data has been entered in column-wise
- Therefore we want to make our 6 condition factors first, where there will be 12 observations per condition (3 pots x 4 plants) per condition
- Then we want to code our nested pots levels. We have 4 plants in each pot. So we want 4 lots of pot 1, then 4 lots of pot 2, then 4 lots of pot 3, per condition, and then repeat that 6 times for all conditions.

```
In [54]: conditions <- factor(c(rep(1:6, rep(12,6))))
```

```
In [55]: pots <- factor(c( rep( rep(1:3, rep(4,3)),6) ))
```

```
In [56]: plants <- data.frame(growth, conditions, pots)

In [57]: rm(growth, conditions, pots)
         attach(plants)
```

2.3 Running the nested ANOVA

```
In [59]: plants.aov <- aov(growth ~ conditions/pots)

         summary(plants.aov)
```

```
                Df Sum Sq Mean Sq F value Pr(>F)
conditions       5 179.64   35.93  38.466 <2e-16 ***
conditions:pots 12  25.83    2.15   2.305 0.0186 *
Residuals       54  50.44    0.93
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

2.4 We're running two hypothesis tests here:

1) $H_{0A} : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = \tau_6 = 0$,

where the alternative hypothesis, H_{1A} is that at least one of the treatment effects for factor A: conditions is non-zero.

2) $H_{0B} : \sigma_B^2 = 0$,

where the alternative hypothesis, H_{1B} is that the variance of random variable representing treatment effect differences between factor B levels (i.e. between pots) **within** factor A (i.e. within each condition) is greater than 0. The intuition of this is that this non-zero variance represents significant differences between the factor B levels within factor A.

2.5 F-tests

1) To test H_{0A}, we construct the following F-statistic: $F_A = \dfrac{MS_A}{MS_{(A)B}}$, which under H_{0A} follows the $F_{a-1,a(b-1)}$ distribution. We can reject the H_{0A} at the $100\alpha\%$ level of significance if $F_A > F_{a-1,a(b-1),\alpha}$. Note this is not the F-value provided in R, as R's default setting is for factors A and B to both be fixed.

2) To test H_{0B}, we construct the following F-statistic: $F_B = \dfrac{MA_{(A)B}}{MS_R}$, which under H_{0B} follows the $F_{a(b-1),ab(n-1)}$ distribution. We can reject the H_{0B} at the $100\alpha\%$ level of significance if $F_B > F_{a(b-1),ab(n-1),\alpha}$. Note this **is** the value provided in R, as this is the common way of testing a treatment effect, by comparing the mean square of the treatment with the mean square of the residuals, representing the pooled estimate of the error variance.

```
In [64]: MS.A <- 35.93
         MS.A.B <- 2.15
         F.A <- MS.A / MS.A.B
         F.A
```

16.7116279069767

```
In [63]: p <- pf(F.A, 5, 12, lower.tail=F)
         p
```

4.84805255930161e-05

2.6 Summary of ANOVA

- We can reject H_{0A} at the 0.1% level of significance, and therefore accept the alternative hypothesis that at least one of the six conditions has a treatement effect that is non-zero.
- We can also reject H_{0B} at the 5% level of significance, and therefore accept the alternative hypothesis that at least one of the pots within the conditions **shows variablity**.