Lab week 10

Two way ANOVA without interactions

In an experiment to investigate the validity of a solution as a liquid absorbance standard, the value of molar absorptivity, ϵ , of solutions of three different concentrations: 0.02, 0.04, 0.06 was calculated at four different wavelengths: 240, 270, 300, 350. One absorbance measurement was made for each combination of concentration and wavelength. The order in which the measurements were made was randomized. The results are shown below.

	Wavelength			
Concentration	240	270	300	350
0.02	91	100	98	100
0.04	93	103	97	98
0.06	95	106	99	99

The two factors: concentration and wavelength may affect the results (the molar absorptivity). We study simultaneously the effects of the two factors by means of two-way ANOVA. We define the wavelength as factor **A** and the concentration as factor **B**. Factor A has 4 levels and factor B has 3. To perform a two-way analysis of variance in R we need to code the table structure. The response values that we model will be all stored in a vector y, while the corresponding rows and columns are contained in two categorical variable A and B. It is important to transform the categorical variables expressed in numbers into factors. If we do not indicate that A and B are factors, we get useless results!

```
y < - c(91,100,98,100, 93,103,97,98, 95,106,99,99)
A < - factor(rep(1:4,3))
A
```

1 2 3 4 1 2 3 4 1 2 3 4

B < - factor(rep(1:3,each=4))

B

1 1 1 1 2 2 2 2 3 3 3 3

We can now execute two-way ANOVA without interactions using the *lm* function.

model <- lm(y ~ A + B)

anova(model)

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	152.25	50.750	19.031	0.001813
B	2	14.00	7.000	2.625	0.151704
Residuals	6	16.00	2.667		

– – –

Two-way ANOVA results are obtained using a sum of squares decomposition. The total sum of squares for a data set is a measure of the variability among all the data,i.e. all molar absorptivity values. The total variability is divided up into three components: variability among the levels of each of the two factors and variability within cells (error variability). Two separate statistical tests are performed (based on the F statistic), comparing each of the first two sources of variability (variability due to first factor and variability due to second factor) to the error variability. In each test, the resulting p-value allows us to determine whether that specific effect is significant.

The first statistical test checks the validity of the null hypothesis

$H_0 : \tau_{240} = \tau_{270} = \tau_{300} = \tau_{350}$

The p-value for the between column factor (between wavelengths) is 0.001813 which is less than 0.05 so factor A is significant. We reject the null hypothesis and accept the alternative that the mean

molar absorptivity depends upon wavelength.

The second statistical test checks the validity of the null hypothesis

$H_0 : \beta_{0.02} = \beta_{0.04} = \beta_{0.06}$

The p-value for the between rows factor (between concentrations) is 0.151704 which is greater than 0.05 so factor B is not significant. We accept the null hypothesis that the mean molar absorptivity is similar for all concentration levels.

The analysis implies that only the main effect of the column factor is significant.

In this approach, the two factors have been assumed to enter the model in a linear fashion (additive factors). We can also include an interaction term in the model. This is done in R by adding the extra term $A : B$ in the parameters of the lm function.

```
model.interactions <- lm(y ~ A + B + A:B)
anova(model.interactions)
Analysis of Variance Table
Response: y
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
A          3  152.25   50.750
B          2   14.00    7.000
A:B        6   16.00    2.667
Residuals  0    0.00
Warning message:
In anova.lm(model.interactions) :
ANOVA F-tests on an essentially perfect fit are unreliable
```

– – –

We see that the SSE becomes zero and the F scores can not be calculated. This is happening because we lack the degrees of freedom to study the interaction effects when we have only one observation

per cell.

Source of variation	**Degrees of freedom**
Total	k*b-1=4*3-1=**11**
Between treatments	k-1=4-1=**3**
Between blocks	b-1=3-1=**2**
Interactions	(k-1)*(b-1)=3*2=**6**

For the following equation to be true we need the residual degrees of freedom, $DF_{Residuals}$, to be zero.

$$\begin{aligned} DF_{total} &= DF_{treatments} + DF_{blocks} + DF_{Interactions} + DF_{Residuals} \\ 11 &= 3 + 2 + 6 + 0 \end{aligned}$$

Zero degrees of freedom for the residuals makes impossible calculating the mean of squared residuals, MSE, which in turn leads to not being able to calculate the F scores. For this reason there is not enough data (degrees of freedom) to test for interactions. In this case we can only study the main factors without interactions.

We can resolve the degrees of freedom problem by replicating the measurements in each cell. So instead of one single measurement in a cell we need more than two. Two-way ANOVA is most powerful when the experiment has the same number of replicates in each group defined by the pair of factors. This is called a balanced design.

Two way ANOVA with interactions

Consider the example of the molar absorptivity from the previous section, but with different results and n=2 measurements for each cell.

	Wavelength			
Concentration	240	270	300	350
0.02	94,96	106,108	48,51	78,81
0.04	93,93	106,105	47,48	78,78
0.06	93,94	106,107	49,50	78,79

To perform a two-way analysis of variance in R we need to code the new table structure.

```
y < - c(94,96,106,108,48,51,78,81,93,93,106,105,47,48,78,78,93,94,106,107,49,50,78,79)

A < - factor(rep(1:4,3,each=2))

A

1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4

B < - factor(rep(1:3,each=8))

B

1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3
```

We can now execute a **balanced two-way ANOVA with interactions** using the *lm* function.

```
model <- lm( y ~ A + B+ A:B)

anova(model)

Analysis of Variance Table

Response: y
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	11059.5	3686.5	2764.875	< 2e-16
B	2	12.3	6.2	4.625	0.03243
A:B	6	2.0	0.3	0.250	0.95000
Residuals	12	16.0	1.3		

– – –

Three separate statistical tests are performed (based on the F statistic), comparing the variability due to first factor, the variability due to second factor and the variability due to the interaction between factors to the error variability.

The error variability is estimated from the within-cell variation and it equals 16 with $(n-1)bk = 1*2*3 = 12$ degrees of freedom. The residual mean square is obtained as the ratio between the residual sum of square and its corresponding degrees of freedom: $MS_{Residuals} = \frac{SS_{Residuals}}{df} = \frac{16}{12} = 1.3$

Each source of variation is compared with the residual mean square to test whether it is significant. The interactions term is very important and before analyzing the main effect we first check if the interaction between the two factors is significant.

Interactions effect. The interaction effect is not significant as the mean sum of squares due to interactions equals 0.3. This values is less then the residuals mean squares, 1.3, which results in a small F test statistic 0.250 and a high p-value 0.95. Since the interaction term is not significant, we can interpret the separate effects of factor A and B.

Between-column effect The effect of factor A is highly significant since we have a very large F score. $F = \frac{3686.5}{1.3} = 2765$ and the p-value is very small making this source of variation very significant.

Between-row effect The effect of factor B is not significant since the p-value 0.03243 is less than 0.05 making this source of variation not significant.