IMM - DTU Statistical Design and Analysis of Experiments
H. Spliid, October 20, 2010

Solution to exercise 31

Except for the missing observations we have a usual two-way analysis of variance design with one observation per cell. It is thus reasonable to apply the following model (without interaction term - why ?)

$$Y_{ij} = \mu + k_i + f_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \in \mathrm{NID}(0, \sigma^2)$$

With the given data this can be written in matrix notation:

$$\begin{Bmatrix} 7.62 \\ 8.02 \\ 7.93 \\ 8.15 \\ 8.12 \\ 7.76 \\ 8.73 \\ 8.74 \\ 8.00 \\ 8.75 \end{Bmatrix} = \begin{Bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{Bmatrix} \begin{Bmatrix} \mu \\ k_1 \\ k_2 \\ k_3 \\ k_4 \\ f_1 \\ f_2 \\ f_3 \end{Bmatrix} + \begin{Bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{31} \\ \varepsilon_{32} \\ \varepsilon_{33} \\ \varepsilon_{41} \\ \varepsilon_{43} \end{Bmatrix}$$

that is in the form: $Y = x\ \beta + \varepsilon$.

The usual linear restrictions for the f's and the k's are

$$\begin{array}{l} \sum r_i k_i = 0; \quad r_i = \text{number of obs. in row no. } i \\ \sum s_j f_j = 0; \quad s_j = \text{number of obs. in column no. } j \end{array}$$

The estimate for β is found by solving the system of equations (the normal equations):

$$[x'x]\ \hat{\beta} = x'\ Y$$

and the linear restrictions.

$$\begin{aligned} \sum r_i k_i &= 0 \\ \sum s_j f_j &= 0 \end{aligned}$$

We can now construct the actual normal equations and solve them:

$$\begin{Bmatrix} 10 & 3 & 2 & 3 & 2 & 3 & 3 & 4 \\ 3 & 3 & 0 & 0 & 0 & 1 & 1 & 1 \\ 2 & 0 & 2 & 0 & 0 & 0 & 1 & 1 \\ 3 & 0 & 0 & 3 & 0 & 1 & 1 & 1 \\ 2 & 0 & 0 & 0 & 2 & 1 & 0 & 1 \\ 3 & 1 & 0 & 1 & 1 & 3 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 0 & 3 & 0 \\ 4 & 1 & 1 & 1 & 1 & 0 & 0 & 4 \end{Bmatrix} \begin{Bmatrix} \mu \\ k_1 \\ k_2 \\ k_3 \\ k_4 \\ f_1 \\ f_2 \\ f_3 \end{Bmatrix} = \begin{Bmatrix} 81.82 \\ 23.57 \\ 16.27 \\ 25.23 \\ 16.75 \\ 23.38 \\ 24.90 \\ 33.54 \end{Bmatrix}$$

The linear restrictions are

$$\begin{Bmatrix} 0 & 3 & 2 & 3 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 4 \end{Bmatrix} \begin{Bmatrix} \mu \\ k_1 \\ k_2 \\ k_3 \\ k_4 \\ f_1 \\ f_2 \\ f_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}$$

Solving these equations gives:

$$\begin{array}{lcllcl} \hat{\mu} &=& 8.1820 & & & \\ \hat{k}_1 &=& -0.3035 & \hat{f}_1 &=& -0.4829 \\ \hat{k}_2 &=& -0.2557 & \hat{f}_2 &=& 0.2211 \\ \hat{k}_3 &=& 0.2498 & \hat{f}_3 &=& 0.1963 \\ \hat{k}_4 &=& 0.3363 & & & \end{array}$$

The deviations between the observations and the corresponding estimated model values are:

	Type of seed		
	A	B_1	B_2
Potash 10	0.2244	-0.0796	-0.1448
20		0.0026	-0.0026
30	-0.1889	0.0770	0.1119
40	-0.0354		0.0354

with sum of squares 0.1343, and it has $10-(8-2)=4$ degrees of freedom. Note that, for example, the sum of deviations is 0 both row-wise and column-wise.

We want test the row effect, that is to determine whether the amount of potash has (significant) influence on the strength index. We will therefore compare the full model above with the following reduced hypothetical (reduced) model, where potash is left out:

$$H_0^{potash}: \; Y_{ij} = \mu + f_j + \varepsilon_{ij}, \qquad \varepsilon_{ij} \in \mathrm{NID}(0, \sigma^2)$$

In the same way as above the general linear model corresponding to H_0 is constructed, and we find the solution:

$$\begin{array}{rcl} \hat{\hat{\mu}} &=& 8.1820 \\ \hat{\hat{f}}_1 &=& -0.3887 \\ \hat{\hat{f}}_2 &=& 0.1180 \\ \hat{\hat{f}}_3 &=& 0.2030 \end{array}$$

and the residual sum of squares is 0.8962, which has a total of $10-(4-1)=7$ degrees of freedom.

The removal of potash from the model led to an increase of the residual sum of squares amounting to $0.8962-0.1343=0.7619$ with $7-4=3$ degrees of freedom. We can now construct the analysis of variance table:

General ANOVA concerning potash			
Variation	SSQ	f	Test quantity
Potash effect	0.7619	3	$z = \frac{0.7619/3}{0.1343/4} = 7.56$
Residual (full model)	0.1343	4	
Deviation from H_0 hypothesis	0.8962	7	

For a test with level of significance 5% the critical value is $z > F(3,4)_{0.05} = 6.59$, wherefore the potash effect is statistically significant at the 5% level.

We now want to assess the type of seed effect in the same way by choosing:

$$H_0^{seed}: \; Y_{ij} = \mu + k_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \in \mathrm{NID}(0, \sigma^2)$$

and we find

$$\begin{aligned}\hat{\hat{\mu}} &= 8.1820\\ \hat{\hat{k}}_1 &= -0.3253\\ \hat{\hat{k}}_2 &= -0.0470\\ \hat{\hat{k}}_3 &= 0.2280\\ \hat{\hat{k}}_4 &= 0.1930\end{aligned}$$

Note that, for example, the linear restriction

$$3\hat{\hat{k}}_1 + 2\hat{\hat{k}}_2 + 3\hat{\hat{k}}_3 + 2\hat{\hat{k}}_4 = 0$$

is satisfied.

General ANOVA concerning type of seed			
Variation	SSQ	f	Test quantity
Seed effect	0.8693	2	$z = \frac{0.8693/2}{0.1343/4} = 12.95$
Residual (full model)	0.1343	4	
Deviation from H_0 hypothesis	1.0036	6	

Since $F(2,4)_{0.05} = 6.94$, the type of seed effect is also statistically significant at the 5% level of significance.

We conclude that both the amount of potash and the type of seed are important for the fiber strength. We therefore stick to the original full model and as estimate for the effects we use the values given on page **??**.

The highest index is obtained by choosing one of B-types of seed.

In praxis the above general procedure is done by applying a proper statistical computer program and the most often used procedure is usually called GLM. It will automatically give the correct F-values and corresponding to the above procedure.

Sometimes is is illustrative to estimate the missing values and analyze the data including these values. The ANOVA then becomes balanced and can be carried out in the usual way provided a correction for degrees of freedom is made.

We show the method:

The missing observations are Y_{21} and Y_{42}. By means of the estimates from the full model we find:

$$\begin{aligned}\hat{Y}_{21} = \hat{\mu} + \hat{k}_2 + \hat{f}_1 &= 8.1820 - 0.2557 - 0.4829\\ &= 7.44\\ \hat{Y}_{42} = \hat{\mu} + \hat{k}_4 + \hat{f}_1 &= 8.1820 + 0.3363 + 0.2211\\ &= 8.74\end{aligned}$$

By adding these data to the original data we get:

	Type of seed			
	A	B_1	B_2	Sum
Potash 0	7.62	8.02	7.93	23.57
10	*7.44*	8.15	8.12	23.71
20	7.76	8.73	8.74	25.23
30	8.00	*8.74*	8.75	25.49
Sum	30.82	33.64	33.54	98.00

And further

$$\begin{aligned}
\text{ssq}_{seed} &= \frac{30.82^2 + 33.64^2 + 33.54^2}{4} - \frac{98.00^2}{12} \\
&= \frac{3206.4536}{4} - 800.3333 = 1.2801 \\
\text{ssq}_{potash} &= \frac{23.57 + 23.71^2 + 25.33^2 + 25.49^2}{3} - \frac{98.00^2}{12} \\
&= 1.007 \\
\text{ssq}_{total} &= 7.62^2 + \cdots + 8.75^2 - \frac{98.00^2}{12} = 2.4151 \\
\text{ssq}_{residual} &= 2.4151 - 1.0007 - 1.2801 = 0.1343
\end{aligned}$$

Since there are only 10 independent observations in the data the ssq_{total}, which is the variation around the total average of the data, has $10 - 1 = 9$ degrees of freedom.

The residual variation thus has $9 - 2 - 3 = 4$ degrees of freedom.

We find as follows:

Approximate ANOVA				
Variation	ssq	f	s^2	Test quantity
Seed	1.2801	2	0.6401	19.05
Potash	1.0007	3	0.3336	9.94
Residual (full model)	0.1343	6-2	0.0336	
Total	2.4151	11-2		

$$F(2,4)_{0.95} = 6,94; \quad F(3,4)_{0.95} = 6.59$$

We conclude that both potash and type of seed are statistically significant as above.

In comparison with the previous analysis both test quantities exhibit a little stronger significance, and that will generally be the case when an approximate analysis is based on estimated values instead what is found from the exact analysis.

Reestimating the model on the data including the estimated values gives:

$$\begin{aligned}
\hat{\mu}^* &= 98/12 = 8.1667 \\
\hat{k}_1^* &= 23.57/3 - 8.1667 = -0.3100 \\
\hat{k}_2^* &= 23.71/3 - 8.1667 = -0.2633 \\
\hat{k}_3^* &= 25.23/3 - 8.1667 = 0.2433 \\
\hat{k}_4^* &= 25.49/3 - 8.1667 = 0.3300 \\
\hat{f}_1^* &= 30.82/4 - 8.1667 = -0.4617 \\
\hat{f}_2^* &= 33.64/4 - 8.1667 = 0.2433 \\
\hat{f}_3^* &= 33.54/4 - 8.1667 = 0.2183
\end{aligned}$$

We (for the sake of illustration) end by splitting the variation corresponding to the effects in the approximate analysis by means of orthogonal contrasts.

With respect to the types of seed it would be reasonable to to see if the A-type deviates from the B-types as such and if the two B-types can be considered not different or not.

We find

$$\begin{array}{lcll} C^{seed}_{A-B} & = & 2T_A - T_{B_1} - T_{B_2} & \\ & = & 2 \cdot 30.82 - 33.64 - 33.54 & = -5.54 \\ C^{seed}_{B_1-B_2} & = & T_{B_1} - T_{B_2} = 33.65 - 33.54 & = 0.10 \\ \text{ssq}_{A-B} & = & (-5.54)^2/(2^2 \cdot 4 + 4 + 4) & = 1.2788 \\ \text{ssq}_{B_1-B_2} & = & 0.1^2/(4+4) & = 0.0013 \\ & & \text{ssq}_{seed} & = 1.2801 \end{array}$$

For potash we apply orthogonal polynomials:

$$\begin{array}{lcll} C^{potash}_{lin.} & = & -3T_0 - 1T_{10} + 1T_{20} + 3T_{30} & \\ & = & 3 \cdot 23.57 - 23.71 + 25.23 + 3 \cdot 25.49 & = 7.28 \\ C^{potash}_{quadr.} & = & -T_0 + T_{10} + T_{20} - T_{30} & = -0.12 \\ C^{potash}_{cubic} & = & -T_0 + 3T_{10} - 3T_{20} + T_{30} & = -2.64 \\ \text{ssq}_{lin.} & = & 7.28^2/(3^2 \cdot 3 + 3 + 3 + 3^2 \cdot 3) & = 0.8833 \\ \text{ssq}_{quadr.} & = & (-0.12)^2/(3+3+3+3) & = 0.0012 \\ \text{ssq}_{cubic} & = & (-2.54)^2/(3 + 3^2 \cdot 3 + 3^2 \cdot 3 + 3) & = 0.1162 \\ & & \text{ssq}_{potash} & = 1.0007 \end{array}$$

Approximate ANOVA with complete splitting of variation

Variation	ssq	f	s^2	Test quantity
A-seed deviates from B-seed	1.2788	1	1.2788	38.06
B_1-seed deviates from B_2-seed	0.0013	1	0.0013	0.04
Linear dep. from potash.	0.8833	1	0.8833	26.29
Quadr. dep. from potash	0.0012	1	0.0012	0.04
Cubic dep. from potash	0.1162	1	0.1162	3.46
Residual	0.1343	4	0.0336	
Total	2.4151	9		

$F(1,4)_{0.95} = 7.71$, and we see that only the A-B effect and the linear potash effect are significant.

The estimate for the change in strength index by using B instead of A is

$$\frac{1}{2}(\hat{f}^*_2 + \hat{f}^*_3) - \hat{f}^*_1 = 0.69$$

For potash we find the linear polynomial:

$$\begin{array}{rcl} u & = & \dfrac{x_{potash} - 15}{10} \\ \lambda_1 & = & 2 \\ A_1 & = & \dfrac{C^{potash}_{linear}}{3 \cdot (3^2 + 1^2 + 1^2 + 3^2)} = \dfrac{7.28}{60} = 0.1213, \end{array}$$

such that

$$\begin{array}{rcl} \text{Index} & = & \text{const} + 0.1213 \cdot 2 \cdot \dfrac{x_{potash} - 15}{10} \\ & = & \text{const} + 0.0243 \cdot x_{potash} \end{array}$$

where x_{potash} = the amount of potash pr kg pr acre.