ORIGIN ≡ 0

W. Stein

Nested 2-Way ANOVA as Linear Models - Unbalanced Example

As with other linear models, unbalanced data require use of the regression approach, in this case by contrast coding of independent variables using a scheme not described in full by KNNL p. 1105. I have been unable to find this scheme in R, so I coded it by hand. Results match KNNL, but do not entirely match R's anova() report using either contr.sum or contr.treatment codings. Example comes from Chapter 26 in Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Example: KNNL dataset Table 26.6

Cell Means ANOVA Model:

< Here the independent variables were explicitly coded into p=4 Dummy variables X1-X4 using KNNL coding scheme designed for nested variable B within A.

K := READPRN("c:/2008LinearModelsData/TrainingSchoolUBcodedR.txt")

Variable Assignment:

$Y := K^{\langle 1\rangle}$

$N := \text{length}(Y) \quad N = 9$ **< total number of cases**

$a := 2 \quad b := 3$

$ii := 0..N-1 \quad i := 0..N-1$

$I := \text{identity}(N) \quad J_{i,ii} := 1$

$OV_i := 1$

$j := 2..5$ **< Identity, One Matrix & One Vector for matrix calculations**

$X^{\langle j-2\rangle} := K^{\langle j\rangle}$

$X := \text{augment}(OV, X)$ **< design matrix**

$r := \text{cols}(X) \quad r = 5 \quad p := r$

$$K = \begin{pmatrix} 1 & 20 & 1 & 1 & 0 & 0 \\ 2 & 22 & 1 & 1 & 0 & 0 \\ 3 & 8 & 1 & 0 & 1 & 0 \\ 4 & 9 & 1 & -1 & -1 & 0 \\ 5 & 13 & 1 & -1 & -1 & 0 \\ 6 & 4 & -1 & 0 & 0 & 1 \\ 7 & 8 & -1 & 0 & 0 & 1 \\ 8 & 16 & -1 & 0 & 0 & -1 \\ 9 & 20 & -1 & 0 & 0 & -1 \end{pmatrix}$$

$$Y = \begin{pmatrix} 20 \\ 22 \\ 8 \\ 9 \\ 13 \\ 4 \\ 8 \\ 16 \\ 20 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & 1 & -1 & -1 & 0 \\ 1 & -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & -1 \end{pmatrix}$$

^ note intercept column in X

Least Squares Estimation of the Regression Parameters:

$$\beta := \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot Y \qquad \beta = \begin{pmatrix} 12.6666667 \\ 0.6666667 \\ 7.6666667 \\ -5.3333333 \\ -6 \end{pmatrix}$$

< meaning of coefficients not determine

Fitted Values & Hat Matrix H:

$Y_h := X \cdot \beta$ **< fitted values Yh**

$H := X \cdot \left(X^T \cdot X\right)^{-1} \cdot X^T$ **< nXn Hat matrix**

Residuals:

$e := Y - Y_h$ **< residuals**

Full Model ANOVA Table for Cell Means Model:

Sum of Squares:		Degrees of Freedom:		Mean Squares:	
$SSTR := Y^T \cdot \left[H - \left(\frac{1}{N}\right) \cdot J\right] \cdot Y$	SSTR = (308)	$df_R := p - 1$	$df_R = 4$	$MSTR := \frac{SSTR}{df_R}$	MSTR = (77)
$SSE := Y^T \cdot (I - H) \cdot Y$	SSE = (26)	$df_E := N - p$	$df_E = 4$	$MSE := \frac{SSE}{df_E}$	MSE = (6.5)
$SSTO := Y^T \cdot \left[I - \left(\frac{1}{N}\right) \cdot J\right] \cdot Y$	SSTO = (334)	$df_T := N - 1$	$df_T = 8$	$MSTO := \frac{SSTO}{df_T}$	MSTO = (41.75)

^ SSTR sums as shown below, SSE verified KNNL p. 1106

GLM Cell Means Decomposition of SSTR:

```
#READ STRUCTURED DATA TABLE WITH NESTED
CODE KNNL p. 1105
K=read.table("c:/2008LinearModelsData/TrainingSc-
hoolUBcodedR.txt")
K
attach(K)
Y=Score
FM=lm(Y~X1+X2+X3+X4)
summary(FM)
anova(FM)
RMa=lm(Y~X2+X3+X4)
RMbina=lm(Y~X1)
anova(RMa,FM)
anova(RMbina,FM)
detach(K)
```

```
> K
  Score X1 X2 X3 X4
1    20  1  1  0  0
2    22  1  1  0  0
3     8  1  0  1  0
4     9  1 -1 -1  0
5    13  1 -1 -1  0
6     4 -1  0  0  1
7     8 -1  0  0  1
8    16 -1  0  0 -1
9    20 -1  0  0 -1
```

```
> summary(FM)
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.6667     0.8760  14.460 0.000133 ***
X1            0.6667     0.8760   0.761 0.489026
X2            7.6667     1.5899   4.822 0.008510 **
X3           -5.3333     1.9003  -2.807 0.048485 *
X4           -6.0000     1.2748  -4.707 0.009262 **
```

```
> anova(FM)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1   12.8    12.8  1.9692 0.233191
X2         1  100.0   100.0 15.3846 0.017214 *
X3         1   51.2    51.2  7.8769 0.048485 *
X4         1  144.0   144.0 22.1538 0.009262 **
Residuals  4   26.0     6.5
```

12.8 + 100 + 51.2 + 144 = 308

^ anova() SS sum to SSTR above

Note: *serial* anova() is appropriate here because X2-X4 need to be considered together, and not as separate marginal estimates.

```
> anova(RMa,FM)
Analysis of Variance Table
Model 1: Y ~ X2 + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      5 29.7647
2      4 26.0000  1    3.7647 0.5792  0.489
```

^ Extra SS for A

```
> anova(RMbina,FM)
Analysis of Variance Table
Model 1: Y ~ X1
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1      7 321.2
2      4  26.0  3     295.2 15.139 0.01195 *
```

^ Extra SS for B in A

$SSA := 3.7647$ **< derived from GLM anova(RM,FM)**

$SS_{BinA} := 295.2$ **^ confirmed KNNL p. 1106**

Prototype in R:

```
#READ STRUCTURED DATA TABLE WITH NUMERIC CODED FACTOR
K=read.table("c:/2008LinearModelsData/TrainingSchoolUBR.txt")
K
attach(K)
Y=Score
A=factor(School)
B=factor(Instructor)
contrasts(A)=contr.sum
contrasts(B)=contr.sum
FM=lm(Y~A+B%in%A)
anova(FM)
detach(K)
```

```
> K
  Score School Instructor
1    20      1          1
2    22      1          1
3     8      1          2
4     9      1          3
5    13      1          3
6     4      2          4
7     8      2          4
8    16      2          5
9    20      2          5
```

```
> anova(FM)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
A          1   12.8    12.8  1.9692 0.23319
A:B        3  295.2    98.4 15.1385 0.01195 *
Residuals  4   26.0     6.5
```

< report for A doesn't match KNNL, but matches summary() above.

Example:

K := READPRN("c:/2008LinearModelsData/TrainingSchoolUBR.txt")

$$K = \begin{pmatrix} 1 & 20 & 1 & 1 \\ 2 & 22 & 1 & 1 \\ 3 & 8 & 1 & 2 \\ 4 & 9 & 1 & 3 \\ 5 & 13 & 1 & 3 \\ 6 & 4 & 2 & 4 \\ 7 & 8 & 2 & 4 \\ 8 & 16 & 2 & 5 \\ 9 & 20 & 2 & 5 \end{pmatrix}$$

Variable Assignment:

$Y := K^{\langle 1 \rangle}$ **< response (dependent) variable Y**

	B_1	B_2	B_3
A_1	$Y_{11} := \begin{pmatrix} 20 \\ 22 \end{pmatrix}$	$Y_{12} := 8$	$Y_{13} := \begin{pmatrix} 9 \\ 13 \end{pmatrix}$
A_2	$Y_{21} := \begin{pmatrix} 4 \\ 8 \end{pmatrix}$	$Y_{22} := \begin{pmatrix} 16 \\ 20 \end{pmatrix}$	

< response blocks

a := 2 b := 3 N := length(Y)

Means & number of elements:

GM := mean(Y) GM = 13.3333 **< grand mean**

$$A := \begin{pmatrix} \text{mean}(20,22,8,9,13) & 5 \\ \text{mean}(4,8,16,20) & 4 \end{pmatrix} \qquad A = \begin{pmatrix} 14.4 & 5 \\ 12 & 4 \end{pmatrix}$$

< Independent Factor A means & n_i

$$B := \begin{pmatrix} \text{mean}(Y_{11}) & 2 & A_{0,0} \\ \text{mean}(Y_{12}) & 1 & A_{0,0} \\ \text{mean}(Y_{13}) & 2 & A_{0,0} \\ \text{mean}(Y_{21}) & 2 & A_{1,0} \\ \text{mean}(Y_{22}) & 2 & A_{1,0} \end{pmatrix} \qquad B = \begin{pmatrix} 21 & 2 & 14.4 \\ 8 & 1 & 14.4 \\ 11 & 2 & 14.4 \\ 6 & 2 & 12 \\ 18 & 2 & 12 \end{pmatrix}$$

< Nested Factor B with A means, n_j & A mean level

$$Y = \begin{pmatrix} 20 \\ 22 \\ 8 \\ 9 \\ 13 \\ 4 \\ 8 \\ 16 \\ 20 \end{pmatrix} \qquad \mu_b := \begin{pmatrix} 21 \\ 21 \\ 8 \\ 11 \\ 11 \\ 6 \\ 6 \\ 18 \\ 18 \end{pmatrix} \qquad W := \text{augment}(Y, \mu_b) \qquad W = \begin{pmatrix} 20 & 21 \\ 22 & 21 \\ 8 & 8 \\ 9 & 11 \\ 13 & 11 \\ 4 & 6 \\ 8 & 6 \\ 16 & 18 \\ 20 & 18 \end{pmatrix}$$

< Within values & associated block means

Sums of Squares:

SSA independent factor:

$i := 0 .. \text{length}\left(A^{\langle 0 \rangle}\right) - 1$

$n := A^{\langle 1 \rangle}$ $AM := A^{\langle 0 \rangle}$

$SSA := \sum_i n_i \cdot \left(AM_i - GM\right)^2$ $SSA = 12.8$

SSB(A) nested factor:

$j := 0 .. \text{length}\left(B^{\langle 0 \rangle}\right) - 1$

$n := B^{\langle 1 \rangle}$ $BM := B^{\langle 0 \rangle}$ $AM := B^{\langle 2 \rangle}$

$SSB := \sum_j n_j \cdot \left(BM_j - AM_j\right)^2$ $SSB = 295.2$

SSE within block error:

$k := 0 .. \text{length}\left(W^{\langle 0 \rangle}\right) - 1$

$O := W^{\langle 0 \rangle}$ $BM := W^{\langle 1 \rangle}$

$SSE := \sum_k \left(O_k - BM_k\right)^2$ $SSE = 26$

Note 1: It remains troubling that I have not found a way in R to automatically recreate ANOVA SS results verified above with KNNL. I presume that different hypotheses $\beta_i = 0$ of regression coefficients derived from the ways R codes nested factor B within A, versus the code system shown in KNNL p. 1105, is responsible for discrepant SSA/MSA/F/P that are reported. The KNNL coding scheme is easily expandable to accomodate any number of factors and/or factor levels, but seemingly this needs to be done by hand. However, having done so, one can use KNNL as the cited reference for how SSA was calculated.

Note 2: The results here conform to the method of calculations in SL Box 10.6. So, Sokal & Rohlf can be cited instead for calculation of Sums of Squares and associated F tests. However, discrepency in hypotheses related to SSA 12.8 here (agreeing with SR) and 3.6747 in above (agreeing with KNNL) remains unresolved.

Note 3: Anova(car) will not work using R's coding for this nested factor. It works fine with the code system in KNNL.

ANOVA Table:

Sums of Squares:		Degrees of freedom:		Mean Squares:	
SSA	$SSA = 12.8$	$df_A := \text{length}\left(A^{\langle 0 \rangle}\right) - 1$	$df_A = 1$	$MSA := \frac{SSA}{df_A}$	$MSA = 12.8$
$SS_{B(A)}$	$SSB = 295.2$	$df_B := \text{length}\left(B^{\langle 0 \rangle}\right) - \text{length}\left(A^{\langle 0 \rangle}\right)$	$df_B = 3$	$MSB := \frac{SSB}{df_B}$	$MSB = 98.4$
SSE	$SSE = 26$	$df_E := \text{length}\left(W^{\langle 0 \rangle}\right) - \text{length}\left(B^{\langle 0 \rangle}\right)$	$df_E = 4$	$MSE := \frac{SSE}{df_E}$	$MSE = 6.5$

Tests of Significance:

Using Full and Reduced Linear Models approach KNNL.

For effect in independent variable A:

Null Hypothesis and Alternative:

H_0: Regression coefficient for Treatment A is zero - no independent effect in A

H_1: Regression coefficient not zero - treatment effect is evident in A

Test Statistic:

$$F_a := \frac{MSA}{MSE} \qquad F_a = 1.96923$$

Decision Rule:

$\alpha := 0.05$ **< set as desired**

If Fs > F(1-α,df_A, df_E) then Reject H_0, otherwise accept H_0

$$qF(1-\alpha, df_A, df_E) = 7.7086$$

Probability:

$$P := \min(pF(F_a, df_A, df_E), 1 - pF(F_a, df_A, df_E)) \qquad P = 0.23319$$

For effect in nested variable B(A):

Null Hypothesis and Alternative:

H_0: Regression coefficient for Nested B is zero - no effect for nested B(A)

H_1: Regression coefficient not zero - effect evident in B(A)

Test Statistic:

$$F_b := \frac{MSB}{MSE} \qquad F_b = 15.13846$$

Decision Rule:

$\alpha := 0.05$ **< set as desired**

If Fs > F(1-α,df_B, df_E) then Reject H_0, otherwise accept H_0

$$qF(1-\alpha, df_B, df_E) = 6.5914$$

Probability:

$$P := \min(pF(F_b, df_B, df_E), 1 - pF(F_b, df_B, df_E)) \qquad P = 0.01195$$

```
> anova(FM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
A          1   12.8    12.8  1.9692 0.23319
A:B        3  295.2    98.4 15.1385 0.01195 *
Residuals  4   26.0     6.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```