

ORIGIN ≡ 0

ANOVA as a Linear Model

W. Stein

ANOVA models are General Linear Models in which discrete (typically unordered) variables defining classes are entered into the Linear Model regression machinery as specially coded "factors". Depending on the coding (i.e., "dummy variables") employed, regression coefficients mean different things. However, the ANOVA tables in ANOVA models are usually the same (one exception noted below). Here, Total Sum of Square SSTO is partitioned into Sum of Squares Treatment SSSTR and Sum of Squares Error SEE, leading to equivalent statistical test results. In many cases, one is content with the ANOVA table results by themselves as summarized (and simplified) by statistical packages such as R. However, it is possible to read meaning from the regression coefficients also. Example comes from Chapter 16 in Kuter et al. (KNNL) *Applied Linear Statistical Models* 5th Edition.

Example:

Kenton Food Exmpl KNNL Table 16.1

Cell Means ANOVA Model: KNNL p.683

< original variable called "Design" has been dummy coded in this dataset as a set of index variables with: p=4 variables for 4 factor levels.

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

Variable Assignment:

```
Y := K<0>
X1 := K<1>    X2 := K<2>    X3 := K<3>    X4 := K<4>

n := length(Y)    n = 19
X := augment(X1, X2, X3, X4)    < design matrix
r := cols(X)    r = 4

p := r
i := 0..n - 1
ii := 0..n - 1

I := identity(n)
Ji,ii := 1
```

< p used previously,
r = p to conform with KNNL

$$K = \begin{pmatrix} 11 & 1 & 0 & 0 & 0 \\ 17 & 1 & 0 & 0 & 0 \\ 16 & 1 & 0 & 0 & 0 \\ 14 & 1 & 0 & 0 & 0 \\ 15 & 1 & 0 & 0 & 0 \\ 12 & 0 & 1 & 0 & 0 \\ 10 & 0 & 1 & 0 & 0 \\ 15 & 0 & 1 & 0 & 0 \\ 19 & 0 & 1 & 0 & 0 \\ 11 & 0 & 1 & 0 & 0 \\ 23 & 0 & 0 & 1 & 0 \\ 20 & 0 & 0 & 1 & 0 \\ 18 & 0 & 0 & 1 & 0 \\ 17 & 0 & 0 & 1 & 0 \\ 27 & 0 & 0 & 0 & 1 \\ 33 & 0 & 0 & 0 & 1 \\ 22 & 0 & 0 & 0 & 1 \\ 26 & 0 & 0 & 0 & 1 \\ 28 & 0 & 0 & 0 & 1 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Least Squares Estimation of the Regression Parameters:

$$b := (X^T \cdot X)^{-1} \cdot X^T \cdot Y \quad b = \begin{pmatrix} 14.6 \\ 13.4 \\ 19.5 \\ 27.2 \end{pmatrix} \quad \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \quad < \text{estimating } \mu \text{ parameters in the model}$$

^ vector of regresion coefficients = cell (block) means

Note: In the Cell Means ANOVA model, regression coefficients in the Linear Model give the estimate of means for each cell (block).

Fitted Values & Hat Matrix H:

$Y_h := X \cdot b$ < fitted values Y_h

$H := X \cdot (X^T \cdot X)^{-1} \cdot X^T$ < $n \times n$ Hat matrix

Residuals:

$e := Y - Y_h$ < residuals

14.6	-3.6
14.6	2.4
14.6	1.4
14.6	-0.6
14.6	0.4
13.4	-1.4
13.4	-3.4
13.4	1.6
13.4	5.6
13.4	-2.4
19.5	3.5
19.5	0.5
19.5	-1.5
19.5	-2.5
27.2	-0.2
27.2	5.8
27.2	-5.2
27.2	-1.2
27.2	0.8

residuals verified KNNL p. 689 ^

ANOVA Table:

Sum of Squares:	Degrees of Freedom:	Mean Squares:
$SSR := Y^T \cdot \left[H - \left(\frac{1}{n} \right) \cdot J \right] \cdot Y$	$SSR = (588.2211)$ $df_R := p - 1$ $df_R = 3$	$MSR := \frac{SSR}{df_R}$ $MSR = (196.0737)$
$SSE := Y^T \cdot (I - H) \cdot Y$	$SSE = (158.2)$ $df_E := n - p$ $df_E = 15$	$MSE := \frac{SSE}{df_E}$ $MSE = (10.5467)$
$SSTO := Y^T \cdot \left[I - \left(\frac{1}{n} \right) \cdot J \right] \cdot Y$	$SSTO = (746.4211)$ $df_T := n - 1$ $df_T = 18$	$MSTO := \frac{SSTO}{df_T}$ $MSTO = (41.4678)$

Note: The ANOVA table is calculated in exactly the same way as for regression. SSR now becomes SSTR (for treatment), both with $df = 3$.

Prototype in R:

```
#CELL MEANS MODEL
#READ STRUCTURED DATA TABLE WITH NUMERIC CODED FACTOR
K=read.table("c:/2008LinearModelsData/KentonFoodCMR.txt")
K
attach(K)
FM=lm(Y~X1+X2+X3+X4-1) #FM for original data K & codes
summary(FM)
model.matrix(FM)
detach(K)
```

< Formula for regression model without initial columns of 1 here

Indicator coded X with p columns = number of levels>

Note: The Cell Means ANOVA Model involves fitting a linear model without the first column of one's. This is a standard option in most computer packages. In R, one specifies this by indicating '-1' as part of the fomula, as above for FM.

> summary(FM)

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 - 1)
Residuals:
    Min       1Q   Median       3Q      Max
-5.20  -1.95  -0.20   1.50   5.80
Coefficients:
    Estimate Std. Error t value Pr(>|t|)
X1    14.600      1.452  10.053 4.66e-08 ***
X2    13.400      1.452   9.226 1.43e-07 ***
X3    19.500      1.624  12.009 4.28e-09 ***
X4    27.200      1.452  18.728 8.16e-12 ***
```

> K						> model.matrix(FM)					
	Y	X1	X2	X3	X4		X1	X2	X3	X4	
1	11	1	0	0	0	1	1	0	0	0	
2	17	1	0	0	0	2	1	0	0	0	
3	16	1	0	0	0	3	1	0	0	0	
4	14	1	0	0	0	4	1	0	0	0	
5	15	1	0	0	0	5	1	0	0	0	
6	12	0	1	0	0	6	0	1	0	0	
7	10	0	1	0	0	7	0	1	0	0	
8	15	0	1	0	0	8	0	1	0	0	
9	19	0	1	0	0	9	0	1	0	0	
10	11	0	1	0	0	10	0	1	0	0	
11	23	0	0	1	0	11	0	0	1	0	
12	20	0	0	1	0	12	0	0	1	0	
13	18	0	0	1	0	13	0	0	1	0	
14	17	0	0	1	0	14	0	0	1	0	
15	27	0	0	0	1	15	0	0	0	1	
16	33	0	0	0	1	16	0	0	0	1	
17	22	0	0	0	1	17	0	0	0	1	
18	26	0	0	0	1	18	0	0	0	1	
19	28	0	0	0	1	19	0	0	0	1	

< Regression coefficients = Cell Means

```
Kc=data.frame(scale(K,center=T,scale=F))
Kc
attach(Kc)
FM=lm(Y~X1+X2+X3+X4-1) #FM for centered data Kc
anova(FM)
detach(Kc)
```

< Note that to obtain the appropriate ANOVA table in R, the data must be mean centered using function scale().

mean-centered data >

```
> Kc
```

	Y	X1	X2	X3	X4
1	-7.6315789	0.7368421	-0.2631579	-0.2105263	-0.2631579
2	-1.6315789	0.7368421	-0.2631579	-0.2105263	-0.2631579
3	-2.6315789	0.7368421	-0.2631579	-0.2105263	-0.2631579
4	-4.6315789	0.7368421	-0.2631579	-0.2105263	-0.2631579
5	-3.6315789	0.7368421	-0.2631579	-0.2105263	-0.2631579
6	-6.6315789	-0.2631579	0.7368421	-0.2105263	-0.2631579
7	-8.6315789	-0.2631579	0.7368421	-0.2105263	-0.2631579
8	-3.6315789	-0.2631579	0.7368421	-0.2105263	-0.2631579
9	0.3684211	-0.2631579	0.7368421	-0.2105263	-0.2631579
10	-7.6315789	-0.2631579	0.7368421	-0.2105263	-0.2631579
11	4.3684211	-0.2631579	-0.2631579	0.7894737	-0.2631579
12	1.3684211	-0.2631579	-0.2631579	0.7894737	-0.2631579
13	-0.6315789	-0.2631579	-0.2631579	0.7894737	-0.2631579
14	-1.6315789	-0.2631579	-0.2631579	0.7894737	-0.2631579
15	8.3684211	-0.2631579	-0.2631579	-0.2105263	0.7368421
16	14.3684211	-0.2631579	-0.2631579	-0.2105263	0.7368421
17	3.3684211	-0.2631579	-0.2631579	-0.2105263	0.7368421
18	7.3684211	-0.2631579	-0.2631579	-0.2105263	0.7368421
19	9.3684211	-0.2631579	-0.2631579	-0.2105263	0.7368421

> anova(FM)

Analysis of Variance Table

		Response: Y				
		Df	Sum Sq	Mean Sq	F value	Pr(>F)
partial SS add to SSTR	>	X1	1 110.29	110.29	11.155	0.004155 **
		X2	1 346.17	346.17	35.011	2.168e-05 ***
		X3	1 131.76	131.76	13.325	0.002157 **
SSE	>	Residuals	16 158.20	9.89		

^ values verified KNNL p. 692

Treatments ANOVA Model as in R:

Note: this the default setting for the function factor() in R. "Treatment Design" involves (p-1)=3 dummy coded variables for 4 factor levels.

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

Variable Assignment:

```
Y := K<0>
X2 := K<2>   X3 := K<3>   X4 := K<4>
n := length(Y)   n = 19
i := 0..n - 1
ii := 0..n - 1
OVi := 1
X := augment(OV, X2, X3, X4) < design matrix
p := cols(X)   p = 4
I := identity(n)
Ji,ii := 1
```

K =	(11 1 0 0 0)	X =	(1 0 0 0)
		17 1 0 0 0				1 0 0 0	
		16 1 0 0 0				1 0 0 0	
		14 1 0 0 0				1 0 0 0	
		15 1 0 0 0				1 0 0 0	
		12 1 1 0 0				1 1 0 0	
		10 1 1 0 0				1 1 0 0	
		15 1 1 0 0				1 1 0 0	
		19 1 1 0 0				1 1 0 0	
		11 1 1 0 0				1 1 0 0	
		23 1 0 1 0				1 0 1 0	
		20 1 0 1 0				1 0 1 0	
		18 1 0 1 0				1 0 1 0	
		17 1 0 1 0				1 0 1 0	
		27 1 0 0 1				1 0 0 1	
		33 1 0 0 1				1 0 0 1	
		22 1 0 0 1				1 0 0 1	
		26 1 0 0 1				1 0 0 1	
		28 1 0 0 1				1 0 0 1	

Least Squares Estimation of the Regression Parameters:

Estimating parameters;

$$b := (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

$$b = \begin{pmatrix} 14.6 \\ -1.2 \\ 4.9 \\ 12.6 \end{pmatrix} \quad \begin{pmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \\ \mu_4 - \mu_1 \end{pmatrix}$$

< cell 1 mean

< cell mean differences from cell mean 1

^ vector of regression coefficients

Note: In the Treatments ANOVA model in R, regression coefficients in the Linear Model give the estimate of means for the first cell (block), followed by differences in means.

Fitted Values & Hat Matrix H:

$$Y_h := X \cdot b \quad < \text{fitted values } Y_h$$

$$H := X \cdot (X^T \cdot X)^{-1} \cdot X^T \quad < \text{nXn Hat matrix}$$

Residuals:

$$e := Y - Y_h \quad < \text{residuals}$$

ANOVA Table:

Sum of Squares:	Degrees of Freedom:	Mean Squares:
$SSR := Y^T \cdot \left[H - \left(\frac{1}{n} \right) \cdot J \right] \cdot Y$	$SSR = (588.2211) \quad df_R := p - 1 \quad df_R = 3$	$MSR := \frac{SSR}{df_R} \quad MSR = (196.0737)$
$SSE := Y^T \cdot (I - H) \cdot Y$	$SSE = (158.2) \quad df_E := n - p \quad df_E = 15$	$MSE := \frac{SSE}{df_E} \quad MSE = (10.5467)$
$SSTO := Y^T \cdot \left[I - \left(\frac{1}{n} \right) \cdot J \right] \cdot Y$	$SSTO = (746.4211) \quad df_T := n - 1 \quad df_T = 18$	$MSTO := \frac{SSTO}{df_T} \quad MSTO = (41.4678)$

Note: The ANOVA table does not change...

Prototype in R:

```
#TREATMENTS MODEL IN R
```

```
#READ STRUCTURED DATA TABLE WITH NUMERIC CODED FACTOR
```

```
K=read.table("c:/2008LinearModelsData/KentonFoodR.txt")
```

```
K
```

```
attach(K)
```

```
Y=Sales
```

```
X=factor(Design) # factor() IN DEFAULT SETTING
```

```
NM=lm(Y~X)
```

```
summary(NM)
```

```
anova(NM)
```

```
model.matrix(NM)
```

```
> summary(NM)
```

```
Call:
```

```
lm(formula = Y ~ X)
```

```
Residuals:
```

```
   Min       1Q   Median       3Q      Max
```

```
 -5.20  -1.95  -0.20   1.50   5.80
```

```
Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  14.600      1.452  10.053 4.66e-08 ***
```

```
X2           -1.200      2.054  -0.584  0.5677
```

```
X3            4.900      2.179   2.249  0.0399 *
```

```
X4           12.600      2.054   6.135 1.91e-05 ***
```

```
> anova(NM)
```

```
Analysis of Variance Table
```

```
Response: Y
```

```
  Df Sum Sq Mean Sq F value    Pr(>F)
```

```
X      3  588.22  196.07  18.591 2.585e-05 ***
```

```
Residuals 15  158.20   10.55
```

```
> K
```

```
Sales Design Store
```

```
1      11      1      1
```

```
2      17      1      2
```

```
3      16      1      3
```

```
4      14      1      4
```

```
5      15      1      5
```

```
6      12      2      1
```

```
7      10      2      2
```

```
8      15      2      3
```

```
9      19      2      4
```

```
10     11      2      5
```

```
11     23      3      1
```

```
12     20      3      2
```

```
13     18      3      3
```

```
14     17      3      4
```

```
15     27      4      1
```

```
16     33      4      2
```

```
17     22      4      3
```

```
18     26      4      4
```

```
19     28      4      5
```

```
> model.matrix(NM)
```

```
(Intercept) X2 X3 X4
```

```
1      1      1  0  0
```

```
2      1      1  0  0
```

```
3      1      1  0  0
```

```
4      1      1  0  0
```

```
5      1      1  0  0
```

```
6      1      1  1  0
```

```
7      1      1  1  0
```

```
8      1      1  1  0
```

```
9      1      1  1  0
```

```
10     1      1  1  0
```

```
11     1      0  1  0
```

```
12     1      0  1  0
```

```
13     1      0  1  0
```

```
14     1      0  1  0
```

```
15     1      0  0  1
```

```
16     1      0  0  1
```

```
17     1      0  0  1
```

```
18     1      0  0  1
```

```
19     1      0  0  1
```

< same ANOVA table

Factor Effects ANOVA Model:

KNNL p. 705

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

Variable Assignment:

```
Y := K<0>
X1 := K<1>    X2 := K<2>    X3 := K<3>
n := length(Y)    n = 19
i := 0..n - 1
ii := 0..n - 1
OVi := 1
X := augment(OV, X1, X2, X3) < design matrix
p := cols(X)    p = 4
I := identity(n)
Ji,ii := 1
```

Note: This "dummy coding" scheme involves a first column of 1's followed by (p-1)=3 coded variables each summing to 0 for 4 variables.

$$K = \begin{pmatrix} 11 & 1 & 0 & 0 \\ 17 & 1 & 0 & 0 \\ 16 & 1 & 0 & 0 \\ 14 & 1 & 0 & 0 \\ 15 & 1 & 0 & 0 \\ 12 & 0 & 1 & 0 \\ 10 & 0 & 1 & 0 \\ 15 & 0 & 1 & 0 \\ 19 & 0 & 1 & 0 \\ 11 & 0 & 1 & 0 \\ 23 & 0 & 0 & 1 \\ 20 & 0 & 0 & 1 \\ 18 & 0 & 0 & 1 \\ 17 & 0 & 0 & 1 \\ 27 & -1 & -1 & -1 \\ 33 & -1 & -1 & -1 \\ 22 & -1 & -1 & -1 \\ 26 & -1 & -1 & -1 \\ 28 & -1 & -1 & -1 \end{pmatrix} \quad X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{pmatrix}$$

Least Squares Estimation of the Regression Parameters:

$$b := (X^T \cdot X)^{-1} \cdot X^T \cdot Y \quad b = \begin{pmatrix} 18.675 \\ -4.075 \\ -5.275 \\ 0.825 \end{pmatrix} \quad \begin{pmatrix} \mu. \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{pmatrix}$$

Estimating parameters:

< Overall mean

< Treatment effect = $\mu. - \mu_i$

$$\tau_4 := -(b_1 + b_2 + b_3) \quad \tau_4 = 8.525 \quad < \tau_4 = -(\tau_1 + \tau_2 + \tau_3)$$

^ vector of regression coefficients, verified KNNL p. 707

In the Factor Effects ANOVA model, the first regression coefficient provides an estimate of the grand mean ($\mu.$) followed by estimates of difference (τ) between grand mean and cell means for the (p-1) blocks. The remaining difference is found by summing the others and then taking the negative.

Fitted Values & Hat Matrix H:

$$Y_h := X \cdot b \quad < \text{fitted values } Y_h$$

$$H := X \cdot (X^T \cdot X)^{-1} \cdot X^T \quad < \text{nXn Hat matrix}$$

Residuals:

$$e := Y - Y_h \quad < \text{residuals}$$

ANOVA Table:

Sum of Squares:	Degrees of Freedom:	Mean Squares:
$SSR := Y^T \cdot \left[H - \left(\frac{1}{n} \right) \cdot J \right] \cdot Y$	$SSR = (588.2211) \quad df_R := p - 1 \quad df_R = 3$	$MSR := \frac{SSR}{df_R} \quad MSR = (196.0737)$
$SSE := Y^T \cdot (I - H) \cdot Y$	$SSE = (158.2) \quad df_E := n - p \quad df_E = 15$	$MSE := \frac{SSE}{df_E} \quad MSE = (10.5467)$
$SSTO := Y^T \cdot \left[I - \left(\frac{1}{n} \right) \cdot J \right] \cdot Y$	$SSTO = (746.4211) \quad df_T := n - 1 \quad df_T = 18$	$MSTO := \frac{SSTO}{df_T} \quad MSTO = (41.4678)$

Note: The ANOVA table does not change...

Prototype in R:

```

#USE OF factor() AS SUM
#READ STRUCTURED DATA TABLE WITH NUMERIC CODED FACTOR
K=read.table("c:/2008LinearModelsData/KentonFoodR.txt")
K
attach(K)
Y=Sales
X=factor(Design)
contrasts(X)=contr.sum #CONVERTS FACTOR
CODING TO SUM

SM=lm(Y~X)
summary(SM)
anova(SM)
model.matrix(SM)

> summary(SM)
Call:
lm(formula = Y ~ X)
Residuals:
    Min       1Q   Median       3Q      Max
-5.20  -1.95  -0.20   1.50   5.80
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.6750    0.7485   24.949 1.25e-13 ***
X1           -4.0750    1.2708   -3.207 0.005884 **
X2           -5.2750    1.2708   -4.151 0.000854 ***
X3            0.8250    1.3706    0.602 0.556221

> anova(SM)
Analysis of Variance Table
Response: Y
            Df Sum Sq Mean Sq F value    Pr(>F)
X              3  588.22   196.07  18.591 2.585e-05 ***
Residuals    15  158.20    10.55

```

< same ANOVA table