Models without an Intercept

Consider the balanced two factor design:

- Factor A: 3 levels, indexed $j = 0, 1, 2$;
- Factor B: 5 levels, indexed $l = 0, 1, 2, 3, 4$;
- $n_{jl} = 4$ replicate observations for each factor level combination;
- total sample size $n = 3 \times 5 \times 4 = 60$.

Data can be simulated as follows:

```
set.seed(2332)
nk<-4; m.A<-3; m.B<-5
n<-nk*m.A*m.B
A<-rep(gl(m.A,m.B,labels=c(0:(m.A-1))),nk)
B<-rep(gl(m.B,1,length=m.A*m.B,labels=c(0:(m.B-1))),nk)
beta.A<-c(0,2,-1)
beta.B<-c(1,1,2,3,1)
beta.AB<-matrix(0,nrow=m.A,ncol=m.B)
beta.AB[2,2:5]<-0.1
beta.AB[3,2:5]<--0.1
Y<-beta.A[as.numeric(A)]+beta.B[as.numeric(B)]+diag(beta.AB[as.numeric(A),as.numeric(B)])+rnorm(n)*2
table(A,B)

:    B
: A   0 1 2 3 4
:   0 4 4 4 4 4
:   1 4 4 4 4 4
:   2 4 4 4 4 4
```

Consider first the model A that fits the intercept plus the A factor. This model supposes that the three groups of 20 subjects, defined by the three levels of factor A, each have a different mean. The mean model is

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \beta_0 + \sum_{j=1}^{2} \beta^{\mathrm{C}}_{Aj} \mathbb{1}_j(A_i) = \begin{cases} \beta_0 & A = 0 \\ \beta_0 + \beta^{\mathrm{C}}_{A1} & A_i = 1 \\ \beta_0 + \beta^{\mathrm{C}}_{A2} & A_i = 2 \end{cases}$$

respectively. We define the **group** means by the usual reparameterization:

$$\beta^{\mathrm{G}}_0 = \beta_0 \qquad \beta^{\mathrm{G}}_1 = \beta_0 + \beta^{\mathrm{C}}_{A1} \qquad \beta^{\mathrm{G}}_2 = \beta_0 + \beta^{\mathrm{C}}_{A2}$$

```
fit1<-lm(Y~A); summary(fit1)

:
: Call:
: lm(formula = Y ~ A)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -4.9040 -1.9531  0.0044  2.0188  5.0067
:
: Coefficients:
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept)   1.9141     0.5354   3.575 0.000722 ***
: A1            1.1201     0.7572   1.479 0.144574
: A2           -1.1159     0.7572  -1.474 0.146068
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 2.394 on 57 degrees of freedom
: Multiple R-squared:  0.1327,Adjusted R-squared:  0.1023
: F-statistic:  4.36 on 2 and 57 DF,  p-value: 0.0173
```

which reveals the estimates $\widehat{\beta}_0 = 1.9141, \widehat{\beta}^{\text{C}}_{A1} = 1.1201, \widehat{\beta}^{\text{C}}_{A2} = -1.1159$ and group mean estimates $\widehat{\beta}^{\text{G}}_0 = 1.9141$,

$$\widehat{\beta}^{\text{G}}_1 = 1.9141 + 1.1201 = 3.0342 \qquad \widehat{\beta}^{\text{G}}_2 = 1.9141 - 1.1159 = 0.7983.$$

Using the formula `Y~ -1+A`, we may reparameterize directly.

```
fit2<-lm(Y~-1+A); summary(fit2)

:
: Call:
: lm(formula = Y ~ -1 + A)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -4.9040 -1.9531  0.0044  2.0188  5.0067
:
: Coefficients:
:    Estimate Std. Error t value Pr(>|t|)
: A0   1.9141     0.5354   3.575 0.000722 ***
: A1   3.0342     0.5354   5.667 5.02e-07 ***
: A2   0.7983     0.5354   1.491 0.141505
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 2.394 on 57 degrees of freedom
: Multiple R-squared:  0.4525,Adjusted R-squared:  0.4237
: F-statistic: 15.71 on 3 and 57 DF,  p-value: 1.462e-07
```

This produces the same estimates of group means $\beta^{\text{G}}_k, k = 0, 1, 2$. The two models yield the same SS_{Res} quantity, as expected.

```
anova(fit1)

: Analysis of Variance Table
:
: Response: Y
:           Df Sum Sq Mean Sq F value Pr(>F)
: A          2  50.00 24.9982    4.36 0.0173 *
: Residuals 57 326.81  5.7336
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(fit2)

: Analysis of Variance Table
:
: Response: Y
:           Df Sum Sq Mean Sq F value    Pr(>F)
: A          3 270.15  90.052  15.706 1.462e-07 ***
: Residuals 57 326.81   5.734
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, the ANOVA tables are **different**, as are the R^2 quantities. This is because the analysis with the intercept **included** uses the decomposition

$$\text{SS}_{\text{T}} = \text{SS}_{\text{Res}} + \text{SS}_{\text{R}}$$

and the model with the intercept **excluded** uses the decomposition

$$\overline{\text{SS}}_{\text{T}} = \text{SS}_{\text{Res}} + \overline{\text{SS}}_{\text{R}}$$

where

$$\text{SS}_{\text{T}} = \sum_{i=1}^{n} y_i^2 - n(\overline{y})^2 = \overline{\text{SS}}_{\text{T}} - n(\overline{y})^2.$$

and $\overline{\text{SS}}_\text{R} = \text{SS}_\text{R} + n(\overline{y})^2$. Thus the difference between the `Sum Sq` quantities is $n\overline{y}^2 = 220.1585031$.

$$n(\overline{y})^2 = 270.15 - 50.00 = 220.15$$

The R^2 values for the models with and without intercepts are

$$\frac{\text{SS}_\text{R}}{\text{SS}_\text{T}} = \frac{\text{SS}_\text{R}}{\text{SS}_\text{Res} + \text{SS}_\text{R}} = \frac{50.00}{50.00 + 326.81} = 0.13269 \quad \text{and} \quad \frac{\overline{\text{SS}}_\text{R}}{\overline{\text{SS}}_\text{T}} = \frac{\overline{\text{SS}}_\text{R}}{\text{SS}_\text{Res} + \overline{\text{SS}}_\text{R}} = \frac{270.15}{270.15 + 326.81} = 0.45254$$

respectively. We now consider the model $A + B$ with and without the intercept:

```
fit3<-lm(Y~A+B);summary(fit3)

:
: Call:
: lm(formula = Y ~ A + B)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -4.0764 -1.7721  0.2442  1.7323  3.8037
:
: Coefficients:
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept)   0.3097     0.7606   0.407  0.68552
: A1            1.1201     0.7042   1.591  0.11764
: A2           -1.1159     0.7042  -1.585  0.11900
: B1            2.1083     0.9091   2.319  0.02428 *
: B2            1.9145     0.9091   2.106  0.03997 *
: B3            3.0233     0.9091   3.326  0.00161 **
: B4            0.9760     0.9091   1.074  0.28789
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 2.227 on 53 degrees of freedom
: Multiple R-squared:  0.3025,Adjusted R-squared:  0.2235
: F-statistic: 3.831 on 6 and 53 DF,  p-value: 0.003026

fit4<-lm(Y~-1+A+B);summary(fit4)

:
: Call:
: lm(formula = Y ~ -1 + A + B)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -4.0764 -1.7721  0.2442  1.7323  3.8037
:
: Coefficients:
:    Estimate Std. Error t value Pr(>|t|)
: A0   0.3097     0.7606   0.407  0.68552
: A1   1.4298     0.7606   1.880  0.06564 .
: A2  -0.8062     0.7606  -1.060  0.29400
: B1   2.1083     0.9091   2.319  0.02428 *
: B2   1.9145     0.9091   2.106  0.03997 *
: B3   3.0233     0.9091   3.326  0.00161 **
: B4   0.9760     0.9091   1.074  0.28789
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 2.227 on 53 degrees of freedom
: Multiple R-squared:  0.5597,Adjusted R-squared:  0.5016
: F-statistic: 9.626 on 7 and 53 DF,  p-value: 1.162e-07
```

The parameter estimates for the levels of factor A change in the same way as before; the contrast parameter estimates do not change.

```
tapply(Y,INDEX=list(A=A,B=B),mean)  #Data means by group

:    B
: A           0        1         2        3          4
:   0  0.4438347 2.918221 2.1054444 3.781890  0.3212851
:   1 -0.3813390 3.287376 3.5894262 4.500478  4.1752655
:   2  0.8708594 1.052791 0.9820351 1.720837 -0.6352464

tapply(fitted(fit3),INDEX=list(A=A,B=B),mean) #Fitted means with intercept

:    B
: A           0        1        2        3         4
:   0  0.3097095 2.418054 2.224226 3.332993 1.2856925
:   1  1.4298158 3.538160 3.344333 4.453099 2.4057988
:   2 -0.8061702 1.302174 1.108347 2.217113 0.1698128

tapply(fitted(fit4),INDEX=list(A=A,B=B),mean) #Fitted means without intercept

:    B
: A           0        1        2        3         4
:   0  0.3097095 2.418054 2.224226 3.332993 1.2856925
:   1  1.4298158 3.538160 3.344333 4.453099 2.4057988
:   2 -0.8061702 1.302174 1.108347 2.217113 0.1698128
```

In an interaction model:

```
fit5<-lm(Y~A*B); summary(fit5)

:
: Call:
: lm(formula = Y ~ A * B)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -4.5253 -1.3724  0.2075  1.3955  4.0841
:
: Coefficients:
:             Estimate Std. Error t value Pr(>|t|)
: (Intercept)   0.4438     1.0951   0.405   0.6872
: A1           -0.8252     1.5488  -0.533   0.5968
: A2            0.4270     1.5488   0.276   0.7840
: B1            2.4744     1.5488   1.598   0.1171
: B2            1.6616     1.5488   1.073   0.2891
: B3            3.3381     1.5488   2.155   0.0365 *
: B4           -0.1225     1.5488  -0.079   0.9373
: A1:B1         1.1943     2.1903   0.545   0.5883
: A2:B1        -2.2925     2.1903  -1.047   0.3009
: A1:B2         2.3092     2.1903   1.054   0.2974
: A2:B2        -1.5504     2.1903  -0.708   0.4827
: A1:B3         1.5438     2.1903   0.705   0.4846
: A2:B3        -2.4881     2.1903  -1.136   0.2620
: A1:B4         4.6792     2.1903   2.136   0.0381 *
: A2:B4        -1.3836     2.1903  -0.632   0.5308
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 2.19 on 45 degrees of freedom
: Multiple R-squared:  0.4271,Adjusted R-squared:  0.2488
: F-statistic: 2.396 on 14 and 45 DF,  p-value: 0.01353
```

```
fit6<-lm(Y~-1+A*B); summary(fit6)

:
: Call:
: lm(formula = Y ~ -1 + A * B)
:
: Residuals:
:     Min      1Q  Median      3Q     Max
: -4.5253 -1.3724  0.2075  1.3955  4.0841
:
: Coefficients:
:       Estimate Std. Error t value Pr(>|t|)
: A0      0.4438     1.0951   0.405   0.6872
: A1     -0.3813     1.0951  -0.348   0.7293
: A2      0.8709     1.0951   0.795   0.4307
: B1      2.4744     1.5488   1.598   0.1171
: B2      1.6616     1.5488   1.073   0.2891
: B3      3.3381     1.5488   2.155   0.0365 *
: B4     -0.1225     1.5488  -0.079   0.9373
: A1:B1   1.1943     2.1903   0.545   0.5883
: A2:B1  -2.2925     2.1903  -1.047   0.3009
: A1:B2   2.3092     2.1903   1.054   0.2974
: A2:B2  -1.5504     2.1903  -0.708   0.4827
: A1:B3   1.5438     2.1903   0.705   0.4846
: A2:B3  -2.4881     2.1903  -1.136   0.2620
: A1:B4   4.6792     2.1903   2.136   0.0381 *
: A2:B4  -1.3836     2.1903  -0.632   0.5308
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
:
: Residual standard error: 2.19 on 45 degrees of freedom
: Multiple R-squared:  0.6384,Adjusted R-squared:  0.5178
: F-statistic: 5.296 on 15 and 45 DF,  p-value: 6.465e-06
```

The fitted values from this model recover the group means exactly:

```
tapply(Y,INDEX=list(A=A,B=B),mean)

:    B
: A           0        1         2        3          4
:   0  0.4438347 2.918221 2.1054444 3.781890  0.3212851
:   1 -0.3813390 3.287376 3.5894262 4.500478  4.1752655
:   2  0.8708594 1.052791 0.9820351 1.720837 -0.6352464

tapply(fitted(fit6),INDEX=list(A=A,B=B),mean)

:    B
: A           0        1         2        3          4
:   0  0.4438347 2.918221 2.1054444 3.781890  0.3212851
:   1 -0.3813390 3.287376 3.5894262 4.500478  4.1752655
:   2  0.8708594 1.052791 0.9820351 1.720837 -0.6352464
```

```
anova(fit3,fit5)

: Analysis of Variance Table
:
: Model 1: Y ~ A + B
: Model 2: Y ~ A * B
:   Res.Df    RSS Df Sum of Sq      F Pr(>F)
: 1     53 262.82
: 2     45 215.88  8    46.941 1.2231 0.3078
```

```r
anova(fit5)

: Analysis of Variance Table
:
: Response: Y
:           Df  Sum Sq Mean Sq F value   Pr(>F)
: A          2  49.996 24.9982  5.2108 0.009215 **
: B          4  63.988 15.9971  3.3345 0.017858 *
: A:B        8  46.941  5.8677  1.2231 0.307797
: Residuals 45 215.883  4.7974
: ---
: Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```