ADJUSTED SUM OF SQUARES

What exactly is the "adjusted sum of squares"?

In the LOWBWT.mtp data, we will consider the RACE and SMOKE factors together. Use subscript i for race; we'll have $i = 1, 2, \ldots, I$. In this case, $I = 3$. We'll use subscript j for smoking; we'll have $j = 1, 2, \ldots, J$. Here we have $J = 2$.

Let n_{ij} be the number of mothers in group (i, j). Using Minitab's **Stat** ⇒ **Tables** ⇒ **Cross tab**, we can get the counts:

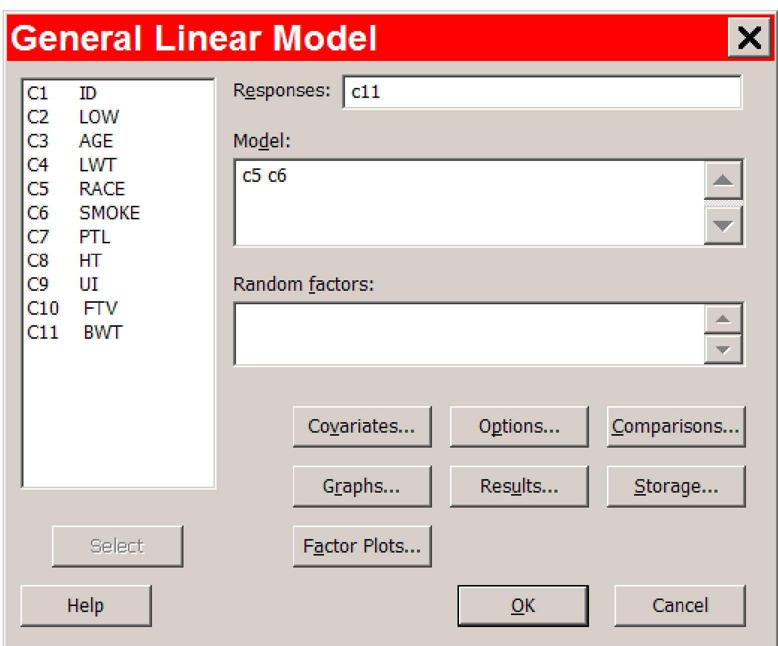
```
Rows: RACE   Columns: SMOKE

          0   1  All

1        44  52   96
2        16  10   26
3        55  12   67
All     115  74  189
```

Subscript k will count within (i, j) combinations. Our model might be

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

The command **Stat** ⇒ **ANOVA** ⇒ **Two Way** requires a balanced design, meaning all n_{ij} have to be the same. We can't use that.

We'll use **Stat** ⇒ **ANOVA** ⇒ **General Linear Model**. The syntax used to specify designs gets complicated. We get this panel:



ADJUSTED SUM OF SQUARES

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

This gets way more complicated, so let's just take this as a start.

General Linear Model: BWT versus RACE, SMOKE

```
Factor  Type   Levels  Values
RACE    fixed       3  1, 2, 3
SMOKE   fixed       2  0, 1


Analysis of Variance for BWT, using Adjusted SS for Tests

Source   DF    Seq SS    Adj SS    Adj MS      F      P
RACE      2   5070608   8768299   4384149   9.26  0.000
SMOKE     1   7271097   7271097   7271097  15.36  0.000
Error   185  87575348  87575348    473380
Total   188  99917053


S = 688.026   R-Sq = 12.35%   R-Sq(adj) = 10.93%


Unusual Observations for BWT

Obs      BWT      Fit  SE Fit  Residual  St Resid
130  4990.00  3335.15   91.75   1654.85      2.43 R
131   709.00  2453.31  122.77  -1744.31     -2.58 R
132  1021.00  3335.15   91.75  -2314.15     -3.39 R
134  1330.00  2880.53   86.29  -1550.53     -2.27 R
135  1474.00  2880.53   86.29  -1406.53     -2.06 R

R denotes an observation with a large standardized residual.
```

We do need some help with the notions of adjusted sums of squares.

As a comparison, let's do this as a regression. We'll need to make an indicator set out of RACE. This uses **Calc ⇒ Make Indicator Variables**. It would be convenient to rename them as RWhite, RBlack, ROther.

Here are critical pieces from that run:

Regression Analysis: BWT versus RBlack, ROther, SMOKE

```
The regression equation is
BWT = 3335 - 451 RBlack - 455 ROther - 427 SMOKE

Predictor     Coef  SE Coef      T      P
Constant   3335.15    91.75  36.35  0.000
RBlack      -451.1    153.1  -2.95  0.004
ROther      -454.6    116.4  -3.90  0.000
SMOKE       -427.2    109.0  -3.92  0.000

S = 688.026   R-Sq = 12.4%   R-Sq(adj) = 10.9%

Analysis of Variance
Source           DF        SS       MS     F      P
Regression        3  12341705  4113902  8.69  0.000
Residual Error  185  87575348   473380
Total           188  99917053

Source  DF   Seq SS
RBlack   1  1525713   [a]
ROther   1  3544895   [a]
SMOKE    1  7271097   [b]
```

ADJUSTED SUM OF SQUARES

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

Now . . . for the very first time, let's look at Seq SS. These are the contributions to SS_{Regr} in the order in which the terms were names in the model. The first two lines sum to 5,070,608. [a] Now let's match up some pieces.

```
Analysis of Variance for BWT, using Adjusted SS for Tests

Source    DF    Seq SS         Adj SS    Adj MS         F      P
RACE       2   5070608  [a]   8768299   4384149      9.26  0.000
SMOKE      1   7271097        7271097   7271097 [b] 15.36  0.000
Error    185  87575348       87575348    473380
Total    188  99917053
```

It's possible now to see what the adjustment is all about. The adjusted sum of squares is the added contribution to SS_{Regr} (or SS_{Model} or SS_{Fit} or … whatever it's called) *after all other effects have been included*.

If the independent variables had been named in the order SMOKE, RBlack, ROther, the Seq SS list would have been this:

```
Source  DF   Seq SS
SMOKE    1  3573406   [c]
RBlack   1  1552316
ROther   1  7215983
```

The last two lines sum to 8,768,299 [d].

We can make this even more definite. The F test that uses the adjusted sum of squares is exactly the same as the partial F test.

That's easy enough to check for this example. If we run the regression of BWT on {RBlack, ROther} we would get this analysis of variance:

```
Analysis of Variance
Source           DF        SS            MS     F      P
Regression        2   5070608 [a]  2535304  4.97  0.008
Residual Error  186  94846445       509927
Total           188  99917053
```

If we run the regression of BWT on {SMOKE} we would get this:

```
Analysis of Variance
Source           DF        SS           MS     F      P
Regression        1   3573406 [c] 3573406  6.94  0.009
Residual Error  187  96343646      515207
Total           188  99917053
```

ADJUSTED SUM OF SQUARES

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

Above we had the regression of BWT on {RBlack, ROther, SMOKE}, and we'll just copy down that analysis of variance, along with the sequential sums of squares.

```
Analysis of Variance
Source           DF        SS       MS     F      P
Regression        3  12341705  4113902  8.69  0.000
Residual Error  185  87575348   473380
Total           188  99917053
```

```
Source DF  Seq SS
RBlack  1 1525713 [a]
ROther  1 3544895 [a]   These sum to 5,070,608
SMOKE   1 7271097 [b]
```

From **Stat** ⇒ **ANOVA** ⇒ **General Linear Model**, we had this:

```
Analysis of Variance for BWT, using Adjusted SS for Tests
Source   DF    Seq SS         Adj SS        Adj MS      F      P
RACE      2   5070608 [a]    8768299 [d]   4384149   9.26  0.000
SMOKE     1   7271097 [b]    7271097 [b]   7271097  15.36  0.000
Error   185  87575348       87575348       473380
Total   188  99917053
```

If we had run **Stat** ⇒ **ANOVA** ⇒ **General Linear Model** with the variables named in the order SMOKE, RACE we would have had this:

```
Analysis of Variance for BWT, using Adjusted SS for Tests
Source   DF    Seq SS       Adj SS        Adj MS      F      P
SMOKE     1   3573406 [c]  7271097 [b]   7271097  15.36  0.000
RACE      2   8768299 [d]  8768299 [d]   4384149   9.26  0.000
Error   185  87575348     87575348        473380
Total   188  99917053
```

The correspondences should now be complete.

The **General Linear Model** method computes the sequential sums of squares, in the order in which the effects were named. It also computes the adjusted sums of squares, accounting for all the other effects. For the last-named effect, these must be the same.

This is equivalent to the use of the partial F test.

These adjusted sums of squares are sometimes called Type III Sums of Squares. The sequential sums of squares are Type I Sums of Squares.

This begs an interesting question ….. what are Type II Sums of Squares? These will become relevant in models with interactions.