# Using Stata for Two-Way Analysis of Variance

We have previously shown how the following two-way ANOVA problem can be solved using SPSS. We will now approach it using Stata.

Problem. A consumer research firm wants to compare three brands of radial tires (X, Y, and Z) in terms of tread life over different road surfaces. Random samples of four tires of each brand are selected for each of three surfaces (asphalt, concrete, gravel). A machine that can simulate road conditions for each of the road surfaces is used to find the tread life (in thousands of miles) of each tire. Construct an ANOVA table and conduct F-tests for the presence of nonzero brand effects, road surface effects, and interaction effects.

| Surface/ Brand | X | Y | Z |
|---|---|---|---|
| Asphalt | 36, 39, 39, 38 | 42, 40, 39, 42 | 32, 36, 35, 34 |
| Concrete | 38, 40, 41, 40 | 42, 45, 48, 47 | 37, 33, 33, 34 |
| Gravel | 34, 32, 34, 35 | 34, 34, 30, 31 | 36, 35, 35, 33 |

Stata Solution. Stata's anova command is pretty straightforward; first you give the DV followed by the IVs. All independent variables are assumed to be categorical unless you explicitly specify otherwise.

```
. anova treadlif surface brand surface*brand

                   Number of obs =      36     R-squared     =  0.8808
                   Root MSE      = 1.72401     Adj R-squared =  0.8454

          Source |  Partial SS    df       MS           F     Prob > F
-----------------+----------------------------------------------------
           Model |  592.722222     8  74.0902778      24.93     0.0000
                 |
         surface |  241.722222     2  120.861111      40.66     0.0000
           brand |  155.388889     2  77.6944444      26.14     0.0000
   surface*brand |  195.611111     4  48.9027778      16.45     0.0000
                 |
        Residual |       80.25    27  2.97222222
-----------------+----------------------------------------------------
           Total |  672.972222    35  19.2277778
```

This is organized a little differently than SPSS's output and does not include a separate entry for the combined main effects, but otherwise the results provided are the same.

In the above, surface*brand represents the interaction of surface and brand. Unlike SPSS, in Stata you must explicitly specify the interaction terms you want included in the model, using the asterisk to combine two or more variables (e.g. a 3-way interaction would look like v1*v2*v3). If you leave the interaction term out, you get

```
. anova treadlif surface brand

                           Number of obs =      36     R-squared     =  0.5901
                           Root MSE      = 2.98308     Adj R-squared =  0.5372

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  397.111111     4  99.2777778      11.16     0.0000
                         |
                 surface |  241.722222     2  120.861111      13.58     0.0001
                   brand |  155.388889     2  77.6944444       8.73     0.0010
                         |
                Residual |  275.861111    31  8.89874552
              -----------+----------------------------------------------------
                   Total |  672.972222    35  19.2277778
```

anova can also present the results as a regression using dummy variables. (Regression using dummy variables will be explained later in the course.)

```
. anova treadlif surface brand surface*brand, regress

      Source |       SS       df       MS              Number of obs =      36
-------------+------------------------------           F(  8,    27) =   24.93
       Model |  592.722222     8  74.0902778           Prob > F      =  0.0000
    Residual |       80.25    27  2.97222222           R-squared     =  0.8808
-------------+------------------------------           Adj R-squared =  0.8454
       Total |  672.972222    35  19.2277778           Root MSE      =   1.724

------------------------------------------------------------------------------
    treadlif        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------------------------------------------------------
_cons                34.75   .8620067    40.31   0.000     32.98131    36.51869
surface
           1          -.5   1.219062    -0.41   0.685    -3.001308    2.001308
           2          -.5   1.219062    -0.41   0.685    -3.001308    2.001308
           3    (dropped)
brand
           1           -1   1.219062    -0.82   0.419    -3.501308    1.501308
           2         -2.5   1.219062    -2.05   0.050    -5.001308    .0013077
           3    (dropped)
surface*brand
        1  1         4.75   1.724013     2.76   0.010     1.212617    8.287383
        1  2            9   1.724013     5.22   0.000     5.462617    12.53738
        1  3    (dropped)
        2  1          6.5   1.724013     3.77   0.001     2.962617    10.03738
        2  2        13.75   1.724013     7.98   0.000     10.21262    17.28738
        2  3    (dropped)
        3  1    (dropped)
        3  2    (dropped)
        3  3    (dropped)
```

The anova command has several other features. You can include continuous variables via use of the continuous parameter. You can estimate complicated designs (typically used in experimental work) that we have not discussed here. If you are doing a regression with a lot of categorical variables, it may be easier to specify it with the anova command rather than regress. Type help anova or read the Stata reference guide if you want more details.