Title

stata.com

anova — Analysis of variance and covariance

Description

The `anova` command fits analysis-of-variance (ANOVA) and analysis-of-covariance (ANCOVA) models for balanced and unbalanced designs, including designs with missing cells; for repeated-measures ANOVA; and for factorial, nested, or mixed designs.

Quick start

One-way ANOVA model of `y` for factor `a`

```
anova y a
```

Two-way full-factorial ANOVA for factors `a` and `b`

```
anova y a b a#b
```

Same as above

```
anova y a##b
```

ANCOVA model including continuous variable `x`

```
anova y a##b c.x
```

Factor `b` nested within `a`

```
anova y a / b|a /
```

Repeated-measures ANOVA with repeated variable `rvar`

```
anova y a rvar, repeated(rvar)
```

Repeated-measures ANOVA with subjects, `idvar`, observed at each level of `rvar`

```
anova y a / idvar|a rvar rvar#a, repeated(rvar)
```

Menu

Statistics > Linear models and related > ANOVA/MANOVA > Analysis of variance and covariance

Syntax

```
anova varname [termlist] [if] [in] [weight] [, options]
```

where *termlist* is a factor-variable list (see [U] **11.4.3 Factor variables**) with the following additional features:

- Variables are assumed to be categorical; use the `c.` factor-variable operator to override this.
- The `|` symbol (indicating nesting) may be used in place of the `#` symbol (indicating interaction).
- The `/` symbol is allowed after a term and indicates that the following term is the error term for the preceding terms.

options	Description
Model	
`repeated(`*varlist*`)`	variables in *term*s that are repeated-measures variables
`partial`	use partial (or marginal) sums of squares
`sequential`	use sequential sums of squares
`noconstant`	suppress constant term
`dropemptycells`	drop empty cells from the design matrix
Adv. model	
`bse(`*term*`)`	between-subjects error term in repeated-measures ANOVA
`bseunit(`*varname*`)`	variable representing lowest unit in the between-subjects error term
`grouping(`*varname*`)`	grouping variable for computing pooled covariance matrix

`bootstrap`, `by`, `fp`, `jackknife`, and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.

Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.

`aweight`s are not allowed with the `jackknife` prefix; see [R] **jackknife**.

`aweight`s and `fweight`s are allowed; see [U] **11.1.6 weight**.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

`repeated(`*varlist*`)` indicates the names of the categorical variables in the *term*s that are to be treated as repeated-measures variables in a repeated-measures ANOVA or ANCOVA.

`partial` presents the ANOVA table using partial (or marginal) sums of squares. This setting is the default. Also see the `sequential` option.

`sequential` presents the ANOVA table using sequential sums of squares.

`noconstant` suppresses the constant term (intercept) from the ANOVA or regression model.

`dropemptycells` drops empty cells from the design matrix. If `c(emptycells)` is set to `keep` (see [R] **set emptycells**), this option temporarily resets it to `drop` before running the ANOVA model. If `c(emptycells)` is already set to `drop`, this option does nothing.

Adv. model

`bse(`*term*`)` indicates the between-subjects error term in a repeated-measures ANOVA. This option is needed only in the rare case when the `anova` command cannot automatically determine the between-subjects error term.

`bseunit(`*varname*`)` indicates the variable representing the lowest unit in the between-subjects error term in a repeated-measures ANOVA. This option is rarely needed because the `anova` command automatically selects the first variable listed in the between-subjects error term as the default for this option.

`grouping(`*varname*`)` indicates a variable that determines which observations are grouped together in computing the covariance matrices that will be pooled and used in a repeated-measures ANOVA. This option is rarely needed because the `anova` command automatically selects the combination of all variables except the first (or as specified in the `bseunit()` option) in the between-subjects error term as the default for grouping observations.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

`anova` uses least squares to fit the linear models known as ANOVA or ANCOVA (henceforth referred to simply as ANOVA models).

If you want to fit one-way ANOVA models, you may find the `oneway` or `loneway` command more convenient; see [R] **oneway** and [R] **loneway**. If you are interested in MANOVA or MANCOVA, see [MV] **manova**.

Structural equation modeling provides a more general framework for fitting ANOVA models; see the *Stata Structural Equation Modeling Reference Manual*.

ANOVA was pioneered by Fisher. It features prominently in his texts on statistical methods and his design of experiments (1925, 1935). Many books discuss ANOVA; see, for instance, Altman (1991); van Belle et al. (2004); Cobb (1998); Snedecor and Cochran (1989); or Winer, Brown, and Michels (1991). For a classic source, see Scheffé (1959). Kennedy and Gentle (1980) discuss ANOVA's computing problems. Edwards (1985) is concerned primarily with the relationship between multiple regression and ANOVA. Acock (2018, chap. 9) illustrates his discussion with Stata output. Repeated-measures ANOVA is discussed in Winer, Brown, and Michels (1991) and Milliken and Johnson (2009). Pioneering work in repeated-measures ANOVA can be found in Box (1954); Geisser and Greenhouse (1958); Huynh and Feldt (1976); and Huynh (1978). For a Stata-specific discussion of ANOVA contrasts, see Mitchell (2012, chap. 7–9; 2015, chap. 4–9).

One-way ANOVA

anova, entered without options, performs and reports standard ANOVA. For instance, to perform a one-way layout of a variable called endog on exog, you would type anova endog exog.

▷ Example 1: One-way ANOVA

We run an experiment varying the amount of fertilizer used in growing apple trees. We test four concentrations, using each concentration in three groves of 12 trees each. Later in the year, we measure the average weight of the fruit.

If all had gone well, we would have had 3 observations on the average weight for each of the four concentrations. Instead, two of the groves were mistakenly leveled by a confused man on a large bulldozer. We are left with the following data:

```
. use http://www.stata-press.com/data/r15/apple
(Apple trees)
. list, abbrev(10) sepby(treatment)
```

	treatment	weight
1.	1	117.5
2.	1	113.8
3.	1	104.4
4.	2	48.9
5.	2	50.4
6.	2	58.9
7.	3	70.4
8.	3	86.9
9.	4	87.7
10.	4	67.3

To obtain one-way ANOVA results, we type

```
. anova weight treatment
```

Number of obs =	10	R-squared =	0.9147	
Root MSE =	9.07002	Adj R-squared =	0.8721	

Source	Partial SS	df	MS	F	Prob>F
Model	5295.5443	3	1765.1814	21.46	0.0013
treatment	5295.5443	3	1765.1814	21.46	0.0013
Residual	493.59167	6	82.265278		
Total	5789.136	9	643.23733		

We find significant (at better than the 1% level) differences among the four concentrations.

Although the output is a usual ANOVA table, let's run through it anyway. Above the table is a summary of the underlying regression. The model was fit on 10 observations, and the root mean squared error (Root MSE) is 9.07. The R^2 for the model is 0.9147, and the adjusted R^2 is 0.8721.

The first line of the table summarizes the model. The sum of squares (Partial SS) for the model is 5295.5 with 3 degrees of freedom (df). This line results in a mean square (MS) of $5295.5/3 \approx 1765.2$.

The corresponding F statistic is 21.46 and has a significance level of 0.0013. Thus the model appears to be significant at the 0.13% level.

The next line summarizes the first (and only) term in the model, `treatment`. Because there is only one term, the line is identical to that for the overall model.

The third line summarizes the residual. The residual sum of squares is 493.59 with 6 degrees of freedom, resulting in a mean squared error of 82.27. The square root of this latter number is reported as the `Root MSE`.

The model plus the residual sum of squares equals the total sum of squares, which is reported as 5789.1 in the last line of the table. This is the total sum of squares of `weight` after removal of the mean. Similarly, the model plus the residual degrees of freedom sum to the total degrees of freedom, 9. Remember that there are 10 observations. Subtracting 1 for the mean, we are left with 9 total degrees of freedom.

◁

❑ Technical note

Rather than using the `anova` command, we could have performed this analysis by using the `oneway` command. Example 1 in [R] **oneway** repeats this same analysis. You may wish to compare the output.

❑

The `regress` command (see [R] **regress**) is used to fit the underlying regression model corresponding to an ANOVA model fit using the `anova` command. Type `regress` after `anova` to see the coefficients, standard errors, etc., of the regression model for the last run of `anova`.

▷ Example 2: Regression table from a one-way ANOVA

Returning to the apple tree experiment, we found that the fertilizer concentration appears to significantly affect the average weight of the fruit. Although that finding is interesting, we next want to know which concentration appears to grow the heaviest fruit. One way to find out is by examining the underlying regression coefficients.

```
. regress, baselevels
```

Source	SS	df	MS	Number of obs	=	10
				F(3, 6)	=	21.46
Model	5295.54433	3	1765.18144	Prob > F	=	0.0013
Residual	493.591667	6	82.2652778	R-squared	=	0.9147
				Adj R-squared	=	0.8721
Total	5789.136	9	643.237333	Root MSE	=	9.07

weight	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
treatment						
1	0	(base)				
2	-59.16667	7.405641	-7.99	0.000	-77.28762	-41.04572
3	-33.25	8.279758	-4.02	0.007	-53.50984	-12.99016
4	-34.4	8.279758	-4.15	0.006	-54.65984	-14.14016
_cons	111.9	5.236579	21.37	0.000	99.08655	124.7134

See [R] **regress** for an explanation of how to read this table. The `baselevels` option of `regress` displays a row indicating the base category for our categorical variable, `treatment`. In summary, we find that concentration 1, the base (omitted) group, produces significantly heavier fruits than concentration 2, 3, and 4; concentration 2 produces the lightest fruits; and concentrations 3 and 4 appear to be roughly equivalent.

◁

▷ Example 3: ANOVA replay

We previously typed `anova weight treatment` to produce and display the ANOVA table for our apple tree experiment. Typing `regress` displays the regression coefficients. We can redisplay the ANOVA table by typing `anova` without arguments:

```
. anova
                  Number of obs =         10    R-squared     =  0.9147
                  Root MSE      =    9.07002    Adj R-squared =  0.8721

          Source | Partial SS         df          MS        F     Prob>F
-----------------+----------------------------------------------------
           Model |  5295.5443          3   1765.1814     21.46  0.0013
                 |
       treatment |  5295.5443          3   1765.1814     21.46  0.0013
                 |
        Residual |  493.59167          6   82.265278
-----------------+----------------------------------------------------
           Total |   5789.136          9   643.23733
```

◁

Two-way ANOVA

You can include multiple explanatory variables with the `anova` command, and you can specify interactions by placing '`#`' between the variable names. For instance, typing `anova y a b` performs a two-way layout of `y` on `a` and `b`. Typing `anova y a b a#b` performs a full two-way factorial layout. The shorthand `anova y a##b` does the same.

With the default partial sums of squares, when you specify interacted terms, the order of the terms does not matter. Typing `anova y a b a#b` is the same as typing `anova y b a b#a`.

▷ Example 4: Two-way factorial ANOVA

The classic two-way factorial ANOVA problem, at least as far as computer manuals are concerned, is a two-way ANOVA design from Afifi and Azen (1979).

Fifty-eight patients, each suffering from one of three different diseases, were randomly assigned to one of four different drug treatments, and the change in their systolic blood pressure was recorded. Here are the data:

	Disease 1	Disease 2	Disease 3
Drug 1	42, 44, 36 13, 19, 22	33, 26, 33 21	31, –3, 25 25, 24
Drug 2	28, 23, 34 42, 13	34, 33, 31 36	3, 26, 28 32, 4, 16
Drug 3	1, 29, 19	11, 9, 7 1, –6	21, 1, 9 3
Drug 4	24, 9, 22 –2, 15	27, 12, 12 –5, 16, 15	22, 7, 25 5, 12

Let's assume that we have entered these data into Stata and stored the data as `systolic.dta`. Below we `use` the data, `list` the first 10 observations, `summarize` the variables, and `tabulate` the control variables:

```
. use http://www.stata-press.com/data/r15/systolic
(Systolic Blood Pressure Data)
. list in 1/10
```

	drug	disease	systolic
1.	1	1	42
2.	1	1	44
3.	1	1	36
4.	1	1	13
5.	1	1	19
6.	1	1	22
7.	1	2	33
8.	1	2	26
9.	1	2	33
10.	1	2	21

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
drug	58	2.5	1.158493	1	4
disease	58	2.017241	.8269873	1	3
systolic	58	18.87931	12.80087	-6	44

```
. tabulate drug disease
```

	Patient's Disease			
Drug Used	1	2	3	Total
1	6	4	5	15
2	5	4	6	15
3	3	5	4	12
4	5	6	5	16
Total	19	19	20	58

Each observation in our data corresponds to one patient, and for each patient we record `drug`, `disease`, and the increase in the systolic blood pressure, `systolic`. The tabulation reveals that the data are not balanced—there are not equal numbers of patients in each `drug`–`disease` cell. Stata does not require that the data be balanced. We can perform a two-way factorial ANOVA by typing

```
. anova systolic drug disease drug#disease
```

Number of obs =	58	R-squared	=	0.4560	
Root MSE =	10.5096	Adj R-squared	=	0.3259	

Source	Partial SS	df	MS	F	Prob>F
Model	4259.3385	11	387.21259	3.51	0.0013
drug	2997.4719	3	999.15729	9.05	0.0001
disease	415.87305	2	207.93652	1.88	0.1637
drug#disease	707.26626	6	117.87771	1.07	0.3958
Residual	5080.8167	46	110.45254		
Total	9340.1552	57	163.86237		

Although Stata's `table` command does not perform ANOVA, it can produce useful summary tables of your data (see [R] **table**):

```
. table drug disease, c(mean systolic) row col f(%8.2f)
```

	Patient's Disease			
Drug Used	1	2	3	Total
1	29.33	28.25	20.40	26.07
2	28.00	33.50	18.17	25.53
3	16.33	4.40	8.50	8.75
4	13.60	12.83	14.20	13.50
Total	22.79	18.21	15.80	18.88

These are simple means and are not influenced by our `anova` model. More useful is the `margins` command (see [R] **margins**) that provides marginal means and adjusted predictions. Because `drug` is the only significant factor in our ANOVA, we now examine the adjusted marginal means for `drug`.

```
. margins drug, asbalanced
Adjusted predictions                            Number of obs     =         58
Expression   : Linear prediction, predict()
at           : drug             (asbalanced)
               disease          (asbalanced)
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
drug						
1	25.99444	2.751008	9.45	0.000	20.45695	31.53194
2	26.55556	2.751008	9.65	0.000	21.01806	32.09305
3	9.744444	3.100558	3.14	0.003	3.503344	15.98554
4	13.54444	2.637123	5.14	0.000	8.236191	18.8527

These adjusted marginal predictions are not equal to the simple drug means (see the total column from the `table` command); they are based upon predictions from our ANOVA model. The `asbalanced` option of `margins` corresponds with the interpretation of the F statistic produced by ANOVA—each cell is given equal weight regardless of its sample size (see the following three technical notes). You can omit the `asbalanced` option and obtain predictive margins that take into account the unequal sample sizes of the cells.

```
. margins drug
Predictive margins                              Number of obs     =         58
Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
drug						
1	25.89799	2.750533	9.42	0.000	20.36145	31.43452
2	26.41092	2.742762	9.63	0.000	20.89003	31.93181
3	9.722989	3.099185	3.14	0.003	3.484652	15.96132
4	13.55575	2.640602	5.13	0.000	8.24049	18.871

◁

❑ Technical note

How do you interpret the significance of terms like `drug` and `disease` in unbalanced data? If you are familiar with SAS, the sums of squares and the F statistic reported by Stata correspond to SAS type III sums of squares. (Stata can also calculate sequential sums of squares, but we will postpone that topic for now.)

Let's think in terms of the following table:

	Disease 1	Disease 2	Disease 3	
Drug 1	μ_{11}	μ_{12}	μ_{13}	$\mu_{1\cdot}$
Drug 2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\cdot}$
Drug 3	μ_{31}	μ_{32}	μ_{33}	$\mu_{3\cdot}$
Drug 4	μ_{41}	μ_{42}	μ_{43}	$\mu_{4\cdot}$
	$\mu_{\cdot 1}$	$\mu_{\cdot 2}$	$\mu_{\cdot 3}$	$\mu_{\cdot\cdot}$

In this table, μ_{ij} is the mean increase in systolic blood pressure associated with drug i and disease j, while $\mu_{i\cdot}$ is the mean for drug i, $\mu_{\cdot j}$ is the mean for disease j, and $\mu_{\cdot\cdot}$ is the overall mean.

If the data are balanced, meaning that there are equal numbers of observations going into the calculation of each mean μ_{ij}, the row means, $\mu_{i\cdot}$, are given by

$$\mu_{i\cdot} = \frac{\mu_{i1} + \mu_{i2} + \mu_{i3}}{3}$$

In our case, the data are not balanced, but we define the $\mu_{i\cdot}$ according to that formula anyway. The test for the main effect of drug is the test that

$$\mu_{1\cdot} = \mu_{2\cdot} = \mu_{3\cdot} = \mu_{4\cdot}$$

To be absolutely clear, the F test of the term `drug`, called the *main effect* of drug, is formally equivalent to the test of the three constraints:

$$\frac{\mu_{11}+\mu_{12}+\mu_{13}}{3} = \frac{\mu_{21}+\mu_{22}+\mu_{23}}{3}$$

$$\frac{\mu_{11}+\mu_{12}+\mu_{13}}{3} = \frac{\mu_{31}+\mu_{32}+\mu_{33}}{3}$$

$$\frac{\mu_{11}+\mu_{12}+\mu_{13}}{3} = \frac{\mu_{41}+\mu_{42}+\mu_{43}}{3}$$

In our data, we obtain a significant F statistic of 9.05 and thus reject those constraints.

❑

❑ Technical note

Stata can display the symbolic form underlying the test statistics it presents, as well as display other test statistics and their symbolic forms; see *Obtaining symbolic forms* in [R] **anova postestimation**. Here is the result of requesting the symbolic form for the main effect of `drug` in our data:

```
. test drug, symbolic
drug
           1  -(r2+r3+r4)
           2   r2
           3   r3
           4   r4
disease
           1  0
           2  0
           3  0
drug#disease
        1  1  -1/3 (r2+r3+r4)
        1  2  -1/3 (r2+r3+r4)
        1  3  -1/3 (r2+r3+r4)
        2  1   1/3 r2
        2  2   1/3 r2
        2  3   1/3 r2
        3  1   1/3 r3
        3  2   1/3 r3
        3  3   1/3 r3
        4  1   1/3 r4
        4  2   1/3 r4
        4  3   1/3 r4
_cons          0
```

This says exactly what we said in the previous technical note.

❑

❑ Technical note

Saying that there is no main effect of a variable is not the same as saying that it has no effect at all. Stata's ability to perform ANOVA on unbalanced data can easily be put to ill use.

For example, consider the following table of the probability of surviving a bout with one of two diseases according to the drug administered to you:

	Disease 1	Disease 2
Drug 1	1	0
Drug 2	0	1

If you have disease 1 and are administered drug 1, you live. If you have disease 2 and are administered drug 2, you live. In all other cases, you die.

This table has no main effects of either drug or disease, although there is a large interaction effect. You might now be tempted to reason that because there is only an interaction effect, you would be indifferent between the two drugs in the absence of knowledge about which disease infects you. Given an equal chance of having either disease, you reason that it does not matter which drug is administered to you—either way, your chances of surviving are 0.5.

You may not, however, have an equal chance of having either disease. If you knew that disease 1 was 100 times more likely to occur in the population, and if you knew that you had one of the two diseases, you would express a strong preference for receiving drug 1.

When you calculate the significance of main effects on unbalanced data, you must ask yourself why the data are unbalanced. If the data are unbalanced for random reasons and you are making predictions for a balanced population, the test of the main effect makes perfect sense. If, however, the data are unbalanced because the underlying populations are unbalanced and you are making predictions for such unbalanced populations, the test of the main effect may be practically—if not statistically—meaningless.

❑

▷ Example 5: ANOVA with missing cells

Stata can perform ANOVA not only on unbalanced populations, but also on populations that are so unbalanced that entire cells are missing. For instance, using our systolic blood pressure data, let's refit the model eliminating the drug 1–disease 1 cell. Because `anova` follows the same syntax as all other Stata commands, we can explicitly specify the data to be used by typing the `if` qualifier at the end of the `anova` command. Here we want to use the data that are not for drug 1 and disease 1:

```
. anova systolic drug##disease if !(drug==1 & disease==1)
                          Number of obs =         52    R-squared     =  0.4545
                          Root MSE      =    10.1615    Adj R-squared =  0.3215
                 Source | Partial SS         df         MS        F     Prob>F
------------------------+----------------------------------------------------
                  Model |   3527.959         10    352.7959      3.42  0.0025
                        |
                   drug |   2686.5783         3   895.52611      8.67  0.0001
                disease |    327.7926         2    163.8963      1.59  0.2168
           drug#disease |    703.0076         5   140.60152      1.36  0.2586
                        |
               Residual |   4233.4833        41   103.25569
------------------------+----------------------------------------------------
                  Total |   7761.4423        51   152.18514
```

Here we used `drug##disease` as a shorthand for `drug disease drug#disease`.

◁

❑ Technical note

The test of the main effect of drug in the presence of missing cells is more complicated than that for unbalanced data. Our underlying tableau now has the following form:

	Disease 1	Disease 2	Disease 3	
Drug 1		μ_{12}	μ_{13}	
Drug 2	μ_{21}	μ_{22}	μ_{23}	$\mu_{2\cdot}$
Drug 3	μ_{31}	μ_{32}	μ_{33}	$\mu_{3\cdot}$
Drug 4	μ_{41}	μ_{42}	μ_{43}	$\mu_{4\cdot}$
		$\mu_{\cdot 2}$	$\mu_{\cdot 3}$	

The hole in the drug 1–disease 1 cell indicates that the mean is unobserved. Considering the main effect of drug, the test is unchanged for the rows in which all the cells are defined:

$$\mu_{2\cdot} = \mu_{3\cdot} = \mu_{4\cdot}$$

The first row, however, requires special attention. Here we want the average outcome for drug 1, which is averaged only over diseases 2 and 3, to be equal to the average values of all other drugs averaged over those same two diseases:

$$\frac{\mu_{12} + \mu_{13}}{2} = \frac{(\mu_{22} + \mu_{23})/2 + (\mu_{32} + \mu_{33})/2 + (\mu_{42} + \mu_{43})/2}{3}$$

Thus the test contains three constraints:

$$\frac{\mu_{21} + \mu_{22} + \mu_{23}}{3} = \frac{\mu_{31} + \mu_{32} + \mu_{33}}{3}$$

$$\frac{\mu_{21} + \mu_{22} + \mu_{23}}{3} = \frac{\mu_{41} + \mu_{42} + \mu_{43}}{3}$$

$$\frac{\mu_{12} + \mu_{13}}{2} = \frac{\mu_{22} + \mu_{23} + \mu_{32} + \mu_{33} + \mu_{42} + \mu_{43}}{6}$$

❑

Stata can calculate two types of sums of squares, *partial* and *sequential*. If you do not specify which sums of squares to calculate, Stata calculates partial sums of squares. The technical notes above have gone into great detail about the definition and use of partial sums of squares. Use the `sequential` option to obtain sequential sums of squares.

❑ Technical note

Before we illustrate sequential sums of squares, consider one more feature of the partial sums. If you know how such things are calculated, you may worry that the terms must be specified in some particular order, that Stata would balk or, even worse, produce different results if you typed, say, `anova drug#disease drug disease` rather than `anova drug disease drug#disease`. We assure you that is not the case.

When you type a model, Stata internally reorganizes the terms, forms the cross-product matrix, inverts it, converts the result to an upper-Hermite form, and then performs the hypothesis tests. As a final touch, Stata reports the results in the same order that you typed the terms.

❑

▷ Example 6: Sequential sums of squares

We wish to estimate the effects on systolic blood pressure of drug and disease by using sequential sums of squares. We want to introduce disease first, then drug, and finally, the interaction of drug and disease:

```
. anova systolic disease drug disease#drug, sequential
                          Number of obs =         58    R-squared     =  0.4560
                          Root MSE      =    10.5096    Adj R-squared =  0.3259
```

Source	Seq. SS	df	MS	F	Prob>F
Model	4259.3385	11	387.21259	3.51	0.0013
disease	488.63938	2	244.31969	2.21	0.1210
drug	3063.4329	3	1021.1443	9.25	0.0001
disease#drug	707.26626	6	117.87771	1.07	0.3958
Residual	5080.8167	46	110.45254		
Total	9340.1552	57	163.86237		

The F statistic on `disease` is now 2.21. When we fit this same model by using partial sums of squares, the statistic was 1.88.

◁

N-way ANOVA

You may include high-order interaction terms, such as a third-order interaction between the variables `A`, `B`, and `C`, by typing `A#B#C`.

▷ Example 7: Three-way factorial ANOVA

We wish to determine the operating conditions that maximize yield for a manufacturing process. There are three temperature settings, two chemical supply companies, and two mixing methods under investigation. Three observations are obtained for each combination of these three factors.

```
. use http://www.stata-press.com/data/r15/manuf
(manufacturing process data)
. describe
Contains data from http://www.stata-press.com/data/r15/manuf.dta
  obs:            36                          manufacturing process data
 vars:             4                          2 Jan 2016 13:28
 size:           144
```

variable name	storage type	display format	value label	variable label
temperature	byte	%9.0g	temp	machine temperature setting
chemical	byte	%9.0g	supplier	chemical supplier
method	byte	%9.0g	meth	mixing method
yield	byte	%9.0g		product yield

```
Sorted by:
```

We wish to perform a three-way factorial ANOVA. We could type

```
. anova yield temp chem temp#chem meth temp#meth chem#meth temp#chem#meth
```

but prefer to use the ## factor-variable operator for brevity.

```
. anova yield temp##chem##meth
```

```
                          Number of obs =         36    R-squared     =  0.5474
                          Root MSE      =    2.62996    Adj R-squared =  0.3399
```

Source	Partial SS	df	MS	F	Prob>F
Model	200.75	11	18.25	2.64	0.0227
temperature	30.5	2	15.25	2.20	0.1321
chemical	12.25	1	12.25	1.77	0.1958
temperature#chemical	24.5	2	12.25	1.77	0.1917
method	42.25	1	42.25	6.11	0.0209
temperature#method	87.5	2	43.75	6.33	0.0062
chemical#method	.25	1	.25	0.04	0.8508
temperature#chemical# method	3.5	2	1.75	0.25	0.7785
Residual	166	24	6.9166667		
Total	366.75	35	10.478571		

The interaction between temperature and method appears to be the important story in these data. A table of means for this interaction is given below.

```
. table method temp, c(mean yield) row col f(%8.2f)
```

mixing method	machine temperature setting low	medium	high	Total
stir	7.50	6.00	6.00	6.50
fold	5.50	9.00	11.50	8.67
Total	6.50	7.50	8.75	7.58

Here our ANOVA is balanced (each cell has the same number of observations), and we obtain the same values as in the table above (but with additional information such as confidence intervals) by using the `margins` command. Because our ANOVA is balanced, using the `asbalanced` option with `margins` would not produce different results. We request the predictive margins for the two terms that appear significant in our ANOVA: `temperature#method` and `method`.

```
. margins temperature#method method
Predictive margins                              Number of obs     =         36
Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
temperature# method						
low#stir	7.5	1.073675	6.99	0.000	5.284044	9.715956
low#fold	5.5	1.073675	5.12	0.000	3.284044	7.715956
medium#stir	6	1.073675	5.59	0.000	3.784044	8.215956
medium#fold	9	1.073675	8.38	0.000	6.784044	11.21596
high#stir	6	1.073675	5.59	0.000	3.784044	8.215956
high#fold	11.5	1.073675	10.71	0.000	9.284044	13.71596
method						
stir	6.5	.6198865	10.49	0.000	5.220617	7.779383
fold	8.666667	.6198865	13.98	0.000	7.387284	9.946049

We decide to use the folding method of mixing and a high temperature in our manufacturing process.

◁

Weighted data

Like all estimation commands, `anova` can produce estimates on weighted data. See [U] **11.1.6 weight** for details on specifying the weight.

▷ Example 8: Three-way factorial ANOVA on grouped data

We wish to investigate the prevalence of byssinosis, a form of pneumoconiosis that can afflict workers exposed to cotton dust. We have data on 5,419 workers in a large cotton mill. We know whether each worker smokes, his or her race, and the dustiness of the work area. The variables are

`smokes`	smoker or nonsmoker in the last five years
`race`	white or other
`workplace`	1 (most dusty), 2 (less dusty), 3 (least dusty)

We wish to fit an ANOVA model explaining the prevalence of byssinosis according to a full factorial model of `smokes`, `race`, and `workplace`.

The data are unbalanced. Moreover, although we have data on 5,419 workers, the data are grouped according to the explanatory variables, along with some other variables, resulting in 72 observations. For each observation, we know the number of workers in the group (`pop`), the prevalence of byssinosis (`prob`), and the values of the three explanatory variables. Thus we wish to fit a three-way factorial model on grouped data.

We begin by showing a bit of the data, which are from Higgins and Koch (1977).

```
. use http://www.stata-press.com/data/r15/byssin
(Byssinosis incidence)

. describe

Contains data from http://www.stata-press.com/data/r15/byssin.dta
  obs:            72                          Byssinosis incidence
 vars:             5                          19 Dec 2016 07:04
 size:           864
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
smokes          int     %8.0g      smokes     Smokes
race            int     %8.0g      race       Race
workplace       int     %8.0g      workplace
                                              Dustiness of workplace
pop             int     %8.0g                 Population size
prob            float   %9.0g                 Prevalence of byssinosis
-------------------------------------------------------------------------------
Sorted by:

. list in 1/5, abbrev(10) divider
```

	smokes	race	workplace	pop	prob
1.	yes	white	most	40	.075
2.	yes	white	less	74	0
3.	yes	white	least	260	.0076923
4.	yes	other	most	164	.152439
5.	yes	other	less	88	0

The first observation in the data represents a group of 40 white workers who smoke and work in a "most" dusty work area. Of those 40 workers, 7.5% have byssinosis. The second observation represents a group of 74 white workers who also smoke but who work in a "less" dusty environment. None of those workers has byssinosis.

Almost every Stata command allows weights. Here we want to weight the data by `pop`. We can, for instance, make a table of the number of workers by their smoking status and race:

```
. tabulate smokes race [fw=pop]
```

	Race		
Smokes	other	white	Total
no	799	1,431	2,230
yes	1,104	2,085	3,189
Total	1,903	3,516	5,419

The `[fw=pop]` at the end of the `tabulate` command tells Stata to count each observation as representing `pop` persons. When making the tally, `tabulate` treats the first observation as representing 40 workers, the second as representing 74 workers, and so on.

Similarly, we can make a table of the dustiness of the workplace:

```
. tabulate workplace [fw=pop]
```

Dustiness of workplace	Freq.	Percent	Cum.
least	3,450	63.66	63.66
less	1,300	23.99	87.65
most	669	12.35	100.00
Total	5,419	100.00	

We can discover the average incidence of byssinosis among these workers by typing

```
. summarize prob [fw=pop]
```

Variable	Obs	Mean	Std. Dev.	Min	Max
prob	5,419	.0304484	.0567373	0	.287037

We discover that 3.04% of these workers have byssinosis. Across all cells, the byssinosis rates vary from 0 to 28.7%. Just to prove that there might be something here, let's obtain the average incidence rates according to the dustiness of the workplace:

```
. table workplace smokes race [fw=pop], c(mean prob)
```

Dustiness of workplace	Race and Smokes			
	other		white	
	no	yes	no	yes
least	.0107527	.0101523	.0081549	.0162774
less	.02	.0081633	.0136612	.0143149
most	.0820896	.1679105	.0833333	.2295082

Let's now fit the ANOVA model.

```
. anova prob workplace smokes race workplace#smokes workplace#race smokes#race
> workplace#smokes#race [aweight=pop]
(sum of wgt is   5.4190e+03)
```

Number of obs =	65	R-squared =	0.8300
Root MSE =	.025902	Adj R-squared =	0.7948

Source	Partial SS	df	MS	F	Prob>F
Model	.17364654	11	.01578605	23.53	0.0000
workplace	.09762518	2	.04881259	72.76	0.0000
smokes	.01303081	1	.01303081	19.42	0.0001
race	.00109472	1	.00109472	1.63	0.2070
workplace#smokes	.01969034	2	.00984517	14.67	0.0000
workplace#race	.00135252	2	.00067626	1.01	0.3718
smokes#race	.00166287	1	.00166287	2.48	0.1214
workplace#smokes#race	.00095084	2	.00047542	0.71	0.4969
Residual	.03555777	53	.0006709		
Total	.2092043	64	.00326882		

Of course, if we want to see the underlying regression, we could type `regress`.

Above we examined simple means of the cells of `workplace#smokes#race`. Our ANOVA shows `workplace`, `smokes`, and their interaction as being the only significant factors in our model. We now examine the predictive marginal mean byssinosis rates for these terms.

```
. margins workplace#smokes workplace smokes
Predictive margins                              Number of obs     =         65
Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
workplace# smokes						
least#no	.0090672	.0062319	1.45	0.152	-.0034323	.0215667
least#yes	.0141264	.0053231	2.65	0.010	.0034497	.0248032
less#no	.0158872	.009941	1.60	0.116	-.0040518	.0358263
less#yes	.0121546	.0087353	1.39	0.170	-.0053662	.0296755
most#no	.0828966	.0182151	4.55	0.000	.0463617	.1194314
most#yes	.2078768	.012426	16.73	0.000	.1829533	.2328003
workplace						
least	.0120701	.0040471	2.98	0.004	.0039526	.0201875
less	.0137273	.0065685	2.09	0.041	.0005526	.0269019
most	.1566225	.0104602	14.97	0.000	.1356419	.177603
smokes						
no	.0196915	.0050298	3.91	0.000	.0096029	.02978
yes	.0358626	.0041949	8.55	0.000	.0274488	.0442765

Smoking combined with the most dusty workplace produces the highest byssinosis rates.

◁

Ronald Aylmer Fisher (1890–1962) (Sir Ronald from 1952) studied mathematics at Cambridge. Even before he finished his studies, he had published on statistics. He worked as a statistician at Rothamsted Experimental Station (1919–1933), as professor of eugenics at University College London (1933–1943), as professor of genetics at Cambridge (1943–1957), and in retirement at the CSIRO Division of Mathematical Statistics in Adelaide. His many fundamental and applied contributions to statistics and genetics mark him as one of the greatest statisticians of all time, including original work on tests of significance, distribution theory, theory of estimation, fiducial inference, and design of experiments.

ANCOVA

You can include multiple explanatory variables with the `anova` command, but unless you explicitly state otherwise by using the `c.` factor-variable operator, all the variables are interpreted as *categorical variables*. Using the `c.` operator, you can designate variables as *continuous* and thus perform ANCOVA.

▷ Example 9: ANCOVA (ANOVA with a continuous covariate)

We have census data recording the death rate (`drate`) and median age (`age`) for each state. The dataset also includes the region of the country in which each state is located (`region`):

```
. use http://www.stata-press.com/data/r15/census2
(1980 Census data by state)
. summarize drate age region
```

Variable	Obs	Mean	Std. Dev.	Min	Max
drate	50	84.3	13.07318	40	107
age	50	29.5	1.752549	24	35
region	50	2.66	1.061574	1	4

`age` is coded in integral years from 24 to 35, and `region` is coded from 1 to 4, with 1 standing for the Northeast, 2 for the North Central, 3 for the South, and 4 for the West.

When we examine the data more closely, we discover large differences in the death rate across regions of the country:

```
. tabulate region, summarize(drate)
```

Census region	Summary of Death rate Mean	Std. Dev.	Freq.
NE	93.444444	7.0553368	9
N Cntrl	88.916667	5.5833899	12
South	88.3125	8.5457104	16
West	68.769231	13.342625	13
Total	84.3	13.073185	50

Naturally, we wonder if these differences might not be explained by differences in the median ages of the populations. To find out, we fit a regression model (via `anova`) of `drate` on `region` and `age`. In the `anova` example below, we treat `age` as a categorical variable.

```
. anova drate region age
```

Number of obs	= 50	R-squared	= 0.7927
Root MSE	= 6.7583	Adj R-squared	= 0.7328

Source	Partial SS	df	MS	F	Prob>F
Model	6638.8653	11	603.53321	13.21	0.0000
region	1320.0097	3	440.00324	9.63	0.0001
age	2237.2494	8	279.65617	6.12	0.0000
Residual	1735.6347	38	45.674598		
Total	8374.5	49	170.90816		

We have the answer to our question: differences in median ages do not eliminate the differences in death rates across the four regions. The ANOVA table summarizes the two terms in the model, `region` and `age`. The `region` term contains 3 degrees of freedom, and the `age` term contains 8 degrees of freedom. Both are significant at better than the 1% level.

The `age` term contains 8 degrees of freedom. Because we did not explicitly indicate that `age` was to be treated as a continuous variable, it was treated as *categorical*, meaning that unique coefficients were estimated for each level of age. The only clue of this labeling is that the number of degrees of freedom associated with the `age` term exceeds 1. The labeling becomes more obvious if we review the regression coefficients:

```
. regress, baselevels
```

Source	SS	df	MS	Number of obs	=	50
				F(11, 38)	=	13.21
Model	6638.86529	11	603.533208	Prob > F	=	0.0000
Residual	1735.63471	38	45.6745977	R-squared	=	0.7927
				Adj R-squared	=	0.7328
Total	8374.5	49	170.908163	Root MSE	=	6.7583

drate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
region						
NE	0	(base)				
N Cntrl	.4428387	3.983664	0.11	0.912	-7.621668	8.507345
South	-.2964637	3.934766	-0.08	0.940	-8.261981	7.669054
West	-13.37147	4.195344	-3.19	0.003	-21.8645	-4.878439
age						
24	0	(base)				
26	-15	9.557677	-1.57	0.125	-34.34851	4.348506
27	14.30833	7.857378	1.82	0.076	-1.598099	30.21476
28	12.66011	7.495513	1.69	0.099	-2.51376	27.83399
29	18.861	7.28918	2.59	0.014	4.104825	33.61717
30	20.87003	7.210148	2.89	0.006	6.273847	35.46621
31	29.91307	8.242741	3.63	0.001	13.22652	46.59963
32	27.02853	8.509432	3.18	0.003	9.802089	44.25498
35	38.925	9.944825	3.91	0.000	18.79275	59.05724
_cons	68.37147	7.95459	8.60	0.000	52.26824	84.47469

The `regress` command displayed the `anova` model as a regression table. We used the `baselevels` option to display the dropped level (or base) for each term.

If we want to treat `age` as a continuous variable, we must prepend `c.` to `age` in our `anova`.

```
. anova drate region c.age
```

Number of obs =	50	R-squared =	0.7203
Root MSE =	7.21483	Adj R-squared =	0.6954

Source	Partial SS	df	MS	F	Prob>F
Model	6032.0825	4	1508.0206	28.97	0.0000
region	1645.6623	3	548.55409	10.54	0.0000
age	1630.4666	1	1630.4666	31.32	0.0000
Residual	2342.4175	45	52.053721		
Total	8374.5	49	170.90816		

The age term now has 1 degree of freedom. The regression coefficients are

```
. regress, baselevels
```

Source	SS	df	MS	Number of obs	=	50
				F(4, 45)	=	28.97
Model	6032.08254	4	1508.02064	Prob > F	=	0.0000
Residual	2342.41746	45	52.0537213	R-squared	=	0.7203
				Adj R-squared	=	0.6954
Total	8374.5	49	170.908163	Root MSE	=	7.2148

drate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
region						
NE	0	(base)				
N Cntrl	1.792526	3.375925	0.53	0.598	-5.006935	8.591988
South	.6979912	3.18154	0.22	0.827	-5.70996	7.105942
West	-13.37578	3.723447	-3.59	0.001	-20.87519	-5.876377
age	3.922947	.7009425	5.60	0.000	2.511177	5.334718
_cons	-28.60281	21.93931	-1.30	0.199	-72.79085	15.58524

Although we started analyzing these data to explain the regional differences in death rate, let's focus on the effect of age for a moment. In our first model, each level of age had a unique death rate associated with it. For instance, the predicted death rate in a north central state with a median age of 28 was

$$0.44 + 12.66 + 68.37 \approx 81.47$$

whereas the predicted death rate from our current model is

$$1.79 + 3.92 \times 28 - 28.60 \approx 82.95$$

Our previous model had an R^2 of 0.7927, whereas our current model has an R^2 of 0.7203. This "small" loss of predictive power accompanies a gain of 7 degrees of freedom, so we suspect that the continuous-age model is as good as the discrete-age model.

◁

❑ Technical note

There is enough information in the two ANOVA tables to attach a statistical significance to our suspicion that the loss of predictive power is offset by the savings in degrees of freedom. Because the continuous-age model is nested within the discrete-age model, we can perform a standard Chow test. For those of us who know such formulas off the top of our heads, the F statistic is

$$\frac{(2342.41746 - 1735.63471)/7}{45.6745977} = 1.90$$

There is, however, a better way.

We can find out whether our continuous model is as good as our discrete model by putting age in the model twice: once as a continuous variable and once as a categorical variable. The categorical variable will then measure deviations around the straight line implied by the continuous variable, and the F test for the significance of the categorical variable will test whether those deviations are jointly zero.

```
. anova drate region c.age age
                           Number of obs =         50    R-squared     =  0.7927
                           Root MSE      =     6.7583    Adj R-squared =  0.7328
                  Source | Partial SS         df          MS         F    Prob>F
-------------------------+----------------------------------------------------
                   Model |  6638.8653         11   603.53321      13.21  0.0000
                         |
                  region |  1320.0097          3   440.00324       9.63  0.0001
                     age |  699.74137          1   699.74137      15.32  0.0004
                     age |  606.78275          7    86.68325       1.90  0.0970
                         |
                Residual |  1735.6347         38   45.674598
-------------------------+----------------------------------------------------
                   Total |     8374.5         49   170.90816
```

We find that the F test for the significance of the (categorical) `age` variable is 1.90, just as we calculated above. It is significant at the 9.7% level. If we hold to a 5% significance level, we cannot reject the null hypothesis that the effect of `age` is linear.

❑

▷ Example 10: Interaction of continuous and categorical variables

In our census data, we still find significant differences across the regions after controlling for the median age of the population. We might now wonder whether the regional differences are differences in level—independent of age—or are instead differences in the regional effects of age. Just as we can interact categorical variables with other categorical variables, we can interact categorical variables with continuous variables.

```
. anova drate region c.age region#c.age
                           Number of obs =         50    R-squared     =  0.7365
                           Root MSE      =    7.24852    Adj R-squared =  0.6926
                  Source | Partial SS         df          MS         F    Prob>F
-------------------------+----------------------------------------------------
                   Model |  6167.7737          7   881.11053      16.77  0.0000
                         |
                  region |   188.7136          3   62.904534       1.20  0.3225
                     age |  873.4256          1    873.4256      16.62  0.0002
              region#age |  135.69116          3   45.230387       0.86  0.4689
                         |
                Residual |  2206.7263         42   52.541102
-------------------------+----------------------------------------------------
                   Total |     8374.5         49   170.90816
```

The `region#c.age` term in our model measures the differences in slopes across the regions. We cannot reject the null hypothesis that there are no such differences. The `region` effect is now "insignificant". This status does not mean that there are no regional differences in death rates because each test is a *marginal* or *partial* test. Here, with `region#c.age` included in the model, `region` is being tested at the point where `age` is zero. Apart from this value not existing in the dataset, it is also a long way from the mean value of `age`, so the test of `region` at this point is meaningless (although it is valid if you acknowledge what is being tested).

To obtain a more sensible test of region, we can subtract the mean from the age variable and use this in the model.

```
. quietly summarize age
. generate mage = age - r(mean)
. anova drate region c.mage region#c.mage
```

```
                   Number of obs =         50    R-squared     =  0.7365
                   Root MSE      =    7.24852    Adj R-squared =  0.6926
```

Source	Partial SS	df	MS	F	Prob>F
Model	6167.7737	7	881.11053	16.77	0.0000
region	1166.1473	3	388.71578	7.40	0.0004
mage	873.4256	1	873.4256	16.62	0.0002
region#mage	135.69116	3	45.230387	0.86	0.4689
Residual	2206.7263	42	52.541102		
Total	8374.5	49	170.90816		

region is significant when tested at the mean of the age variable.

◁

Remember that we can specify interactions by typing *varname*#*varname*. We have seen examples of interacting categorical variables with categorical variables and, in the examples above, a categorical variable (region) with a continuous variable (age or mage).

We can also interact continuous variables with continuous variables. To include an $\texttt{age}^2$ term in our model, we could type c.age#c.age. If we also wanted to interact the categorical variable region with the $\texttt{age}^2$ term, we could type region#c.age#c.age (or even c.age#region#c.age).

Nested designs

In addition to specifying interaction terms, nested terms can also be specified in an ANOVA. A vertical bar is used to indicate nesting: A|B is read as A nested within B. A|B|C is read as A nested within B, which is nested within C. A|B#C is read as A is nested within the interaction of B and C. A#B|C is read as the interaction of A and B, which is nested within C.

Different error terms can be specified for different parts of the model. The forward slash is used to indicate that the next term in the model is the error term for what precedes it. For instance, anova y A / B|A indicates that the F test for A is to be tested by using the mean square from B|A in the denominator. Error terms (terms following the slash) are generally not tested unless they are themselves followed by a slash. Residual error is the default error term.

For example, consider *A* / *B* / *C*, where *A*, *B*, and *C* may be arbitrarily complex terms. Then anova will report *A* tested by *B* and *B* tested by *C*. If we add one more slash on the end to form *A* / *B* / *C* /, then anova will also report *C* tested by the residual error.

▷ Example 11: Simple nested ANOVA

We have collected data from a manufacturer that is evaluating which of five different brands of machinery to buy to perform a particular function in an assembly line. Twenty assembly-line employees were selected at random for training on these machines, with four employees assigned to learn a particular machine. The output from each employee (operator) on the brand of machine for which he trained was measured during four trial periods. In this example, the operator is nested

within machine. Because of sickness and employee resignations, the final data are not balanced. The following table gives the mean output and sample size for each machine and operator combination.

```
. use http://www.stata-press.com/data/r15/machine, clear
(machine data)
. table machine operator, c(mean output n output) col f(%8.2f)
```

five brands of machine	operator nested in machine 1	2	3	4	Total
1	9.15	9.48	8.27	8.20	8.75
	2	4	3	4	13
2	15.03	11.55	11.45	11.52	12.47
	3	2	2	4	11
3	11.27	10.13	11.13		10.84
	3	3	3		9
4	16.10	18.97	15.35	16.60	16.65
	3	3	4	3	13
5	15.30	14.35	10.43		13.63
	4	4	3		11

Assuming that `operator` is random (that is, we wish to infer to the larger population of possible operators) and `machine` is fixed (that is, only these five machines are of interest), the typical test for `machine` uses `operator` nested within `machine` as the error term. `operator` nested within `machine` can be tested by residual error. Our earlier warning concerning designs with either unplanned missing cells or unbalanced cell sizes, or both, also applies to interpreting the ANOVA results from this unbalanced nested example.

```
. anova output machine / operator|machine /
                          Number of obs =         57    R-squared     =  0.8661
                          Root MSE      =    1.47089    Adj R-squared =  0.8077
```

Source	Partial SS	df	MS	F	Prob>F
Model	545.82229	17	32.107193	14.84	0.0000
machine	430.98079	4	107.7452	13.82	0.0001
operator|machine	101.3538	13	7.7964465		
operator|machine	101.3538	13	7.7964465	3.60	0.0009
Residual	84.376658	39	2.1635041		
Total	630.19895	56	11.253553		

`operator|machine` is preceded by a slash, indicating that it is the error term for the terms before it (here `machine`). `operator|machine` is also followed by a slash that indicates it should be tested with residual error. The output lists the `operator|machine` term twice, once as the error term for machine, and again as a term tested by residual error. A line is placed in the ANOVA table to separate the two. In general, a dividing line is placed in the output to separate the terms into groups that are tested with the same error term. The overall model is tested by residual error and is separated from the rest of the table by a blank line at the top of the table.

The results indicate that the machines are not all equal and that there are significant differences between operators.

◁

▷ Example 12: ANOVA with multiple levels of nesting

Your company builds and operates sewage treatment facilities. You want to compare two particulate solutions during the particulate reduction step of the sewage treatment process. For each solution, two area managers are randomly selected to implement and oversee the change to the new treatment process in two of their randomly chosen facilities. Two workers at each of these facilities are trained to operate the new process. A measure of particulate reduction is recorded at various times during the month at each facility for each worker. The data are described below.

```
. use http://www.stata-press.com/data/r15/sewage
(Sewage treatment)
. describe
Contains data from http://www.stata-press.com/data/r15/sewage.dta
  obs:            64                          Sewage treatment
 vars:             5                          9 May 2016 12:43
 size:           320
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
particulate     byte    %9.0g                 particulate reduction
solution        byte    %9.0g                 2 particulate solutions
manager         byte    %9.0g                 2 managers per solution
facility        byte    %9.0g                 2 facilities per manager
worker          byte    %9.0g                 2 workers per facility
-------------------------------------------------------------------------------
Sorted by: solution  manager  facility  worker
```

You want to determine if the two particulate solutions provide significantly different particulate reduction. You would also like to know if `manager`, `facility`, and `worker` are significant effects. `solution` is a fixed factor, whereas `manager`, `facility`, and `worker` are random factors.

In the following `anova` command, we use abbreviations for the variable names, which can sometimes make long ANOVA model statements easier to read.

```
. anova particulate s / m|s / f|m|s / w|f|m|s /, dropemptycells
                          Number of obs =         64    R-squared     =  0.6338
                          Root MSE      =    12.7445    Adj R-squared =  0.5194
```

Source	Partial SS	df	MS	F	Prob>F
Model	13493.609	15	899.57396	5.54	0.0000
solution	7203.7656	1	7203.7656	17.19	0.0536
manager|solution	838.28125	2	419.14063		
manager|solution	838.28125	2	419.14063	0.55	0.6166
facility|manager|solution	3064.9375	4	766.23438		
facility|manager|solution	3064.9375	4	766.23438	2.57	0.1193
worker|facility|manager|solution	2386.625	8	298.32813		
worker|facility|manager|solution	2386.625	8	298.32813	1.84	0.0931
Residual	7796.25	48	162.42188		
Total	21289.859	63	337.93428		

While `solution` is not declared significant at the 5% significance level, it is near enough to that threshold to warrant further investigation (see example 3 in [R] **anova postestimation** for a continuation of the analysis of these data).

◁

❑ Technical note

Why did we use the `dropemptycells` option with the previous `anova`? By default, Stata retains empty cells when building the design matrix and currently treats `|` and `#` the same in how it determines the possible number of cells. Retaining empty cells in an ANOVA with nested terms can cause your design matrix to become too large. In example 12, there are $1024 = 2 \times 4 \times 8 \times 16$ cells that are considered possible for the `worker|facility|manager|solution` term because the `worker`, `facility`, and `manager` variables are uniquely numbered. With the `dropemptycells` option, the `worker|facility|manager|solution` term requires just 16 columns in the design matrix (corresponding to the 16 unique workers).

Why did we not use the `dropemptycells` option in example 11, where `operator` is nested in `machine`? If you look at the table presented at the beginning of that example, you will see that `operator` is compactly instead of uniquely numbered (you need both `operator` number and `machine` number to determine the `operator`). Here the `dropemptycells` option would have only reduced our design matrix from 26 columns down to 24 columns (because there were only 3 operators instead of 4 for `machine`s 3 and 5).

We suggest that you specify `dropemptycells` when there are nested terms in your ANOVA. You could also use the `set emptycells drop` command to accomplish the same thing; see [R] **set**.

❑

Mixed designs

An ANOVA can consist of both nested and crossed terms. A split-plot ANOVA design provides an example.

▷ Example 13: Split-plot ANOVA

Two reading programs and three skill-enhancement techniques are under investigation. Ten classes of first-grade students were randomly assigned so that five classes were taught with one reading program and another five classes were taught with the other. The 30 students in each class were divided into six groups with 5 students each. Within each class, the six groups were divided randomly so that each of the three skill-enhancement techniques was taught to two of the groups within each class. At the end of the school year, a reading assessment test was administered to all the students. In this split-plot ANOVA, the whole-plot treatment is the two reading programs, and the split-plot treatment is the three skill-enhancement techniques.

```
. use http://www.stata-press.com/data/r15/reading
(Reading experiment data)
. describe
Contains data from http://www.stata-press.com/data/r15/reading.dta
  obs:           300                          Reading experiment data
 vars:             5                          9 Mar 2016 18:57
 size:         1,500                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
score           byte    %9.0g                 reading score
program         byte    %9.0g                 reading program
class           byte    %9.0g                 class nested in program
skill           byte    %9.0g                 skill enhancement technique
group           byte    %9.0g                 group nested in class and skill
-------------------------------------------------------------------------------
Sorted by:
```

In this split-plot ANOVA, the error term for `program` is `class` nested within `program`. The error term for `skill` and the `program` by `skill` interaction is the `class` by `skill` interaction nested within `program`. Other terms are also involved in the model and can be seen below.

Our `anova` command is too long to fit on one line of this manual. Where we have chosen to break the command into multiple lines is arbitrary. If we were typing this command into Stata, we would just type along and let Stata automatically wrap across lines, as necessary.

```
. anova score prog / class|prog skill prog#skill / class#skill|prog /
> group|class#skill|prog /, dropemptycells
```

```
                        Number of obs =        300    R-squared     =  0.3738
                        Root MSE      =    14.6268    Adj R-squared =  0.2199
```

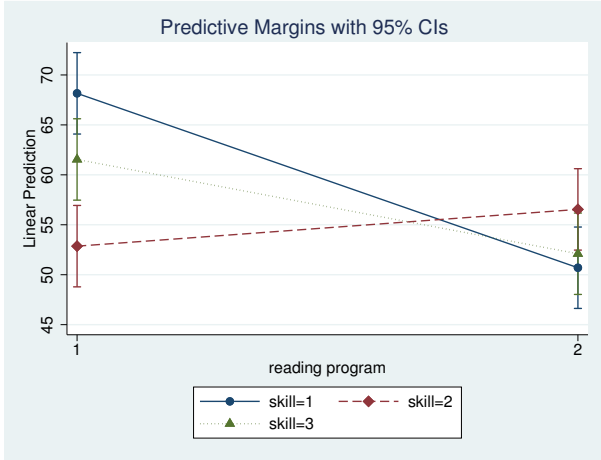
Source	Partial SS	df	MS	F	Prob>F
Model	30656.517	59	519.60198	2.43	0.0000
program	4493.07	1	4493.07	8.73	0.0183
class|program	4116.6133	8	514.57667		
skill	1122.6467	2	561.32333	1.54	0.2450
program#skill	5694.62	2	2847.31	7.80	0.0043
class#skill|program	5841.4667	16	365.09167		
class#skill|program	5841.4667	16	365.09167	1.17	0.3463
group|class#skill| program	9388.1	30	312.93667		
group|class#skill| program	9388.1	30	312.93667	1.46	0.0636
Residual	51346.4	240	213.94333		
Total	82002.917	299	274.25725		

The `program#skill` term is significant, as is the `program` term. Let's look at the predictive margins for these two terms and at a `marginsplot` for the first term.

```
. margins, within(program skill)
Predictive margins                                Number of obs     =        300
Expression   : Linear prediction, predict()
within       : program skill
Empty cells  : reweight
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
program#skill						
1 1	68.16	2.068542	32.95	0.000	64.08518	72.23482
1 2	52.86	2.068542	25.55	0.000	48.78518	56.93482
1 3	61.54	2.068542	29.75	0.000	57.46518	65.61482
2 1	50.7	2.068542	24.51	0.000	46.62518	54.77482
2 2	56.54	2.068542	27.33	0.000	52.46518	60.61482
2 3	52.1	2.068542	25.19	0.000	48.02518	56.17482

```
. marginsplot, plot2opts(lp(dash) m(D)) plot3opts(lp(dot) m(T))
  Variables that uniquely identify margins: program skill
```




```
. margins, within(program)
Predictive margins                                Number of obs     =        300
Expression   : Linear prediction, predict()
within       : program
Empty cells  : reweight
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
program						
1	60.85333	1.194273	50.95	0.000	58.50074	63.20593
2	53.11333	1.194273	44.47	0.000	50.76074	55.46593

Because our ANOVA involves nested terms, we used the `within()` option of `margins`; see [R] **margins**.

`skill` 2 produces a low score when combined with `program` 1 and a high score when combined with `program` 2, demonstrating the interaction between the reading program and the skill-enhancement technique. You might conclude that the first reading program and the first skill-enhancement technique perform best when combined. However, notice the overlapping confidence interval for the first reading program and the third skill-enhancement technique.

◁

❑ Technical note

There are several valid ways to write complicated `anova` terms. In the reading experiment example (example 13), we had a term `group|class#skill|program`. This term can be read as `group` nested within both `class` and `skill` and further nested within `program`. You can also write this term as `group|class#skill#program` or `group|program#class#skill` or `group|skill#class|program`, etc. All variations will produce the same result. Some people prefer having only one '`|`' in a term and would use `group|class#skill#program`, which is read as `group` nested within `class`, `skill`, and `program`.

❑

Gertrude Mary Cox (1900–1978) was born on a farm near Dayton, Iowa. Initially intending to become superintendent of an orphanage, she enrolled at Iowa State College. There she majored in mathematics and attained the college's first Master's degree in statistics. After working on her PhD in psychological statistics for two years at the University of California–Berkeley, she decided to go back to Iowa State to work with George W. Snedecor. There she pursued her interest in and taught a course in design of experiments. That work led to her collaboration with W. G. Cochran, which produced a classic text. In 1940, when Snedecor shared with her his list of men he was nominating to head the statistics department at North Carolina State College, she wanted to know why she had not been included. He added her name, she won the position, and she built an outstanding department at North Carolina State. Cox retired early so she could work at the Research Triangle Institute in North Carolina. She consulted widely, served as editor of *Biometrics*, and was elected to the National Academy of Sciences.

Latin-square designs

You can use `anova` to analyze a Latin-square design. Consider the following example, published in Snedecor and Cochran (1989).

▷ Example 14: Latin-square ANOVA

Data from a Latin-square design are as follows:

Row	Column 1	Column 2	Column 3	Column 4	Column 5
1	257(B)	230(E)	279(A)	287(C)	202(D)
2	245(D)	283(A)	245(E)	280(B)	260(C)
3	182(E)	252(B)	280(C)	246(D)	250(A)
4	203(A)	204(C)	227(D)	193(E)	259(B)
5	231(C)	271(D)	266(B)	334(A)	338(E)

In Stata, the data might appear as follows:

```
. use http://www.stata-press.com/data/r15/latinsq
. list
```

	row	c1	c2	c3	c4	c5
1.	1	257	230	279	287	202
2.	2	245	283	245	280	260
3.	3	182	252	280	246	250
4.	4	203	204	227	193	259
5.	5	231	271	266	334	338

Before `anova` can be used on these data, the data must be organized so that the outcome measurement is in one column. `reshape` is inadequate for this task because there is information about the treatments in the sequence of these observations. `pkshape` is designed to reshape this type of data; see [R] **pkshape**.

```
. pkshape row row c1-c5, order(beacd daebc ebcda acdeb cdbae)
. list
```

	sequence	outcome	treat	carry	period
1.	1	257	1	0	1
2.	2	245	5	0	1
3.	3	182	2	0	1
4.	4	203	3	0	1
5.	5	231	4	0	1
6.	1	230	2	1	2
7.	2	283	3	5	2
8.	3	252	1	2	2
9.	4	204	4	3	2
10.	5	271	5	4	2
11.	1	279	3	2	3
12.	2	245	2	3	3
13.	3	280	4	1	3
14.	4	227	5	4	3
15.	5	266	1	5	3
16.	1	287	4	3	4
17.	2	280	1	2	4
18.	3	246	5	4	4
19.	4	193	2	5	4
20.	5	334	3	1	4
21.	1	202	5	4	5
22.	2	260	4	1	5
23.	3	250	3	5	5
24.	4	259	1	2	5
25.	5	338	2	3	5

```
. anova outcome sequence period treat
```

```
                  Number of obs =         25    R-squared     =  0.6536
                  Root MSE      =    32.4901    Adj R-squared =  0.3073
```

Source	Partial SS	df	MS	F	Prob>F
Model	23904.08	12	1992.0067	1.89	0.1426
sequence	13601.36	4	3400.34	3.22	0.0516
period	6146.16	4	1536.54	1.46	0.2758
treat	4156.56	4	1039.14	0.98	0.4523
Residual	12667.28	12	1055.6067		
Total	36571.36	24	1523.8067		

◁

These methods will work with any type of Latin-square design, including those with replicated measurements. For more information, see [R] **pk**, [R] **pkcross**, and [R] **pkshape**.

Repeated-measures ANOVA

One approach for analyzing repeated-measures data is to use multivariate ANOVA (MANOVA); see [MV] **manova**. In this approach, the data are placed in wide form (see [D] **reshape**), and the repeated measures enter the MANOVA as dependent variables.

A second approach for analyzing repeated measures is to use `anova`. However, one of the underlying assumptions for the F tests in ANOVA is independence of observations. In a repeated-measures design, this assumption is almost certainly violated. In a repeated-measures ANOVA, the subjects (or whatever the experimental units are called) are observed for each level of one or more of the other categorical variables in the model. These variables are called the repeated-measure variables. Observations from the same subject are likely to be correlated, though this is only a problem if the observations violate compound symmetry or the sphericity condition.

The approach used in repeated-measures ANOVA to correct for violation of compound symmetry or sphericity is to apply correction to the degrees of freedom of the F test for terms in the model that involve repeated measures. This correction factor, ϵ, lies between the reciprocal of the degrees of freedom for the repeated term and 1. Box (1954) provided the pioneering work in this area. Milliken and Johnson (2009) refer to the lower bound of this correction factor as Box's conservative correction factor. Winer, Brown, and Michels (1991) call it simply the conservative correction factor.

Geisser and Greenhouse (1958) provide an estimate for the correction factor called the Greenhouse–Geisser ϵ. This value is estimated from the data. Huynh and Feldt (1976) show that the Greenhouse–Geisser ϵ tends to be conservatively biased. They provide a revised correction factor called the Huynh–Feldt ϵ. When the Huynh–Feldt ϵ exceeds 1, it is set to 1. Thus there is a natural ordering for these correction factors:

$$\text{Box's conservative } \epsilon \leq \text{Greenhouse–Geisser } \epsilon \leq \text{Huynh–Feldt } \epsilon \leq 1$$

A correction factor of 1 is the same as no correction.

`anova` with the `repeated()` option computes these correction factors and displays the revised test results in a table that follows the standard ANOVA table. In the resulting table, H-F stands for Huynh–Feldt, G-G stands for Greenhouse–Geisser, and Box stands for Box's conservative ϵ.

▷ Example 15: Repeated-measures ANOVA

This example is taken from table 4.3 of Winer, Brown, and Michels (1991). The reaction time for five subjects each tested with four drugs was recorded in the variable `score`. Here is a table of the data (see [P] **tabdisp** if you are unfamiliar with `tabdisp`):

```
. use http://www.stata-press.com/data/r15/t43, clear
(T4.3 -- Winer, Brown, Michels)
. tabdisp person drug, cellvar(score)
```

person	drug 1	2	3	4
1	30	28	16	34
2	14	18	10	22
3	24	20	18	30
4	38	34	20	44
5	26	28	14	30

`drug` is the repeated variable in this simple repeated-measures ANOVA example. The ANOVA is specified as follows:

```
. anova score person drug, repeated(drug)
                          Number of obs =         20    R-squared     =  0.9244
                          Root MSE      =    3.06594    Adj R-squared =  0.8803

                 Source | Partial SS         df          MS         F    Prob>F
------------------------+----------------------------------------------------
                  Model |       1379          7         197     20.96  0.0000
                        |
                 person |      680.8          4       170.2     18.11  0.0001
                   drug |      698.2          3  232.73333     24.76  0.0000
                        |
                Residual |      112.8         12         9.4
------------------------+----------------------------------------------------
                  Total |     1491.8         19  78.515789

Between-subjects error term:  person
                      Levels:  5          (4 df)
          Lowest b.s.e. variable:  person
Repeated variable: drug
                                        Huynh-Feldt epsilon        =  1.0789
                                        *Huynh-Feldt epsilon reset to 1.0000
                                        Greenhouse-Geisser epsilon =  0.6049
                                        Box's conservative epsilon =  0.3333

                                                  ------------ Prob > F ------------
                 Source |  df       F        Regular     H-F      G-G        Box
------------------------+----------------------------------------------------------
                   drug |   3    24.76      0.0000    0.0000   0.0006     0.0076
                Residual |  12
------------------------+----------------------------------------------------------
```

Here the Huynh–Feldt ϵ is 1.0789, which is larger than 1. It is reset to 1, which is the same as making no adjustment to the standard test computed in the main ANOVA table. The Greenhouse–Geisser ϵ is 0.6049, and its associated p-value is computed from an F ratio of 24.76 using 1.8147 ($= 3\epsilon$) and 7.2588 ($= 12\epsilon$) degrees of freedom. Box's conservative ϵ is set equal to the reciprocal of the degrees of freedom for the repeated term. Here it is 1/3, so Box's conservative test is computed using 1 and 4 degrees of freedom for the observed F ratio of 24.76.

Even for Box's conservative ϵ, `drug` is significant with a p-value of 0.0076. The following table gives the predictive marginal mean `score` (that is, response time) for each of the four drugs:

```
. margins drug
Predictive margins                              Number of obs     =         20
Expression   : Linear prediction, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |
           1 |       26.4   1.371131    19.25   0.000     23.41256    29.38744
           2 |       25.6   1.371131    18.67   0.000     22.61256    28.58744
           3 |       15.6   1.371131    11.38   0.000     12.61256    18.58744
           4 |         32   1.371131    23.34   0.000     29.01256    34.98744
------------------------------------------------------------------------------
```

The ANOVA table for this example provides an F test for `person`, but you should ignore it. An appropriate test for `person` would require replication (that is, multiple measurements for `person` and `drug` combinations). Also, without replication there is no test available for investigating the interaction between `person` and `drug`.

◁

▷ Example 16: Repeated-measures ANOVA with nesting

Table 7.7 of Winer, Brown, and Michels (1991) provides another repeated-measures ANOVA example. There are four dial shapes and two methods for calibrating dials. Subjects are nested within calibration method, and an accuracy score is obtained. The data are shown below.

```
. use http://www.stata-press.com/data/r15/t77
(T7.7 -- Winer, Brown, Michels)
. tabdisp shape subject calib, cell(score)
```

4 dial shapes	2 methods for calibrating dials and subject nested in calib: 1			2		
	1	2	3	1	2	3
1	0	3	4	4	5	7
2	0	1	3	2	4	5
3	5	5	6	7	6	8
4	3	4	2	8	6	9

The calibration method and dial shapes are fixed factors, whereas subjects are random. The appropriate test for calibration method uses the nested subject term as the error term. Both the dial shape and the interaction between dial shape and calibration method are tested with the dial shape by subject interaction nested within calibration method. Here we drop this term from the `anova` command, and it becomes residual error. The dial shape is the repeated variable because each subject is tested with all four dial shapes. Here is the `anova` command that produces the desired results:

```
. anova score calib / subject|calib shape calib#shape, repeated(shape)
```

```
                    Number of obs =         24    R-squared     =  0.8925
                    Root MSE      =    1.11181    Adj R-squared =  0.7939
```

Source	Partial SS	df	MS	F	Prob>F
Model	123.125	11	11.193182	9.06	0.0003
calib	51.041667	1	51.041667	11.89	0.0261
subject|calib	17.166667	4	4.2916667		
shape	47.458333	3	15.819444	12.80	0.0005
calib#shape	7.4583333	3	2.4861111	2.01	0.1662
Residual	14.833333	12	1.2361111		
Total	137.95833	23	5.9981884		

```
Between-subjects error term:  subject|calib
                     Levels:  6         (4 df)
     Lowest b.s.e. variable:  subject
     Covariance pooled over:  calib     (for repeated variable)
Repeated variable: shape
                                  Huynh-Feldt epsilon        =  0.8483
                                  Greenhouse-Geisser epsilon =  0.4751
                                  Box's conservative epsilon =  0.3333
```

Source	df	F	Prob > F: Regular	H-F	G-G	Box
shape	3	12.80	0.0005	0.0011	0.0099	0.0232
calib#shape	3	2.01	0.1662	0.1791	0.2152	0.2291
Residual	12					

The repeated-measure ϵ corrections are applied to any terms that are tested in the main ANOVA table and have the repeated variable in the term. These ϵ corrections are given in a table below the main ANOVA table. Here the repeated-measures tests for `shape` and `calib#shape` are presented.

Calibration method is significant, as is dial shape. The interaction between calibration method and dial shape is not significant. The repeated-measure ϵ corrections do not change these conclusions, but they do change the significance level for the tests on `shape` and `calib#shape`. Here, though, unlike in the example 15, the Huynh–Feldt ϵ is less than 1.

Here are the predictive marginal mean scores for calibration method and dial shapes. Because the interaction was not significant, we request only the `calib` and `shape` predictive margins.

```
. margins, within(calib)
Predictive margins                              Number of obs     =         24
Expression   : Linear prediction, predict()
within       : calib
Empty cells  : reweight
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
calib						
1	3	.3209506	9.35	0.000	2.300709	3.699291
2	5.916667	.3209506	18.43	0.000	5.217375	6.615958

```
. margins, within(shape)
Predictive margins                              Number of obs     =         24
Expression   : Linear prediction, predict()
within       : shape
Empty cells  : reweight
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
shape						
1	3.833333	.4538926	8.45	0.000	2.844386	4.82228
2	2.5	.4538926	5.51	0.000	1.511053	3.488947
3	6.166667	.4538926	13.59	0.000	5.17772	7.155614
4	5.333333	.4538926	11.75	0.000	4.344386	6.32228

◁

❑ Technical note

The computation of the Greenhouse–Geisser and Huynh–Feldt epsilons in a repeated-measures ANOVA requires the number of levels and degrees of freedom for the between-subjects error term, as well as a value computed from a pooled covariance matrix. The observations are grouped based on all but the lowest-level variable in the between-subjects error term. The covariance over the repeated variables is computed for each resulting group, and then these covariance matrices are pooled. The dimension of the pooled covariance matrix is the number of levels of the repeated variable (or combination of levels for multiple repeated variables). In example 16, there are four levels of the repeated variable (`shape`), so the resulting covariance matrix is 4×4.

The `anova` command automatically attempts to determine the between-subjects error term and the lowest-level variable in the between-subjects error term to group the observations for computation of the pooled covariance matrix. `anova` issues an error message indicating that the `bse()` or `bseunit()` option is required when `anova` cannot determine them. You may override the default selections of

anova by specifying the bse(), bseunit(), or grouping() option. The term specified in the bse() option must be a term in the ANOVA model.

The default selection for the between-subjects error term (the bse() option) is the interaction of the nonrepeated categorical variables in the ANOVA model. The first variable listed in the between-subjects error term is automatically selected as the lowest-level variable in the between-subjects error term but can be overridden with the bseunit(*varname*) option. *varname* is often a term, such as subject or subsample within subject, and is most often listed first in the term because of the nesting notation of ANOVA. This term makes sense in most repeated-measures ANOVA designs when the terms of the model are written in standard form. For instance, in example 16, there were three categorical variables (subject, calib, and shape), with shape being the repeated variable. Here anova looked for a term involving only subject and calib to determine the between-subjects error term. It found subject|calib as the term with six levels and 4 degrees of freedom. anova then picked subject as the default for the bseunit() option (the lowest variable in the between-subjects error term) because it was listed first in the term.

The grouping of observations proceeds, based on the different combinations of values of the variables in the between-subjects error term, excluding the lowest level variable (as found by default or as specified with the bseunit() option). You may specify the grouping() option to change the default grouping used in computing the pooled covariance matrix.

The between-subjects error term, number of levels, degrees of freedom, lowest variable in the term, and grouping information are presented after the main ANOVA table and before the rest of the repeated-measures output.

❑

▷ Example 17: Repeated-measures ANOVA with two repeated variables

Data with two repeated variables are given in table 7.13 of Winer, Brown, and Michels (1991). The accuracy scores of subjects making adjustments to three dials during three different periods are recorded. Three subjects are exposed to a certain noise background level, whereas a different set of three subjects is exposed to a different noise background level. Here is a table of accuracy scores for the noise, subject, period, and dial variables:

```
. use http://www.stata-press.com/data/r15/t713
(T7.13 -- Winer, Brown, Michels)
. tabdisp subject dial period, by(noise) cell(score) stubwidth(11)
```

noise background and subject nested in noise	10 minute time periods and type of dial								
	1			2			3		
	1	2	3	1	2	3	1	2	3
1									
1	45	53	60	40	52	57	28	37	46
2	35	41	50	30	37	47	25	32	41
3	60	65	75	58	54	70	40	47	50
2									
1	50	48	61	25	34	51	16	23	35
2	42	45	55	30	37	43	22	27	37
3	56	60	77	40	39	57	31	29	46

noise, period, and dial are fixed, whereas subject is random. Both period and dial are repeated variables. The ANOVA for this example is specified next.

```
. anova score noise / subject|noise period noise#period /
> period#subject|noise dial noise#dial /
> dial#subject|noise period#dial noise#period#dial, repeated(period dial)
```

```
                    Number of obs =         54    R-squared     =  0.9872
                    Root MSE      =    2.81859    Adj R-squared =  0.9576
```

Source	Partial SS	df	MS	F	Prob>F
Model	9797.7222	37	264.8033	33.33	0.0000
noise	468.16667	1	468.16667	0.75	0.4348
subject|noise	2491.1111	4	622.77778		
period	3722.3333	2	1861.1667	63.39	0.0000
noise#period	333	2	166.5	5.67	0.0293
period#subject|noise	234.88889	8	29.361111		
dial	2370.3333	2	1185.1667	89.82	0.0000
noise#dial	50.333333	2	25.166667	1.91	0.2102
dial#subject|noise	105.55556	8	13.194444		
period#dial	10.666667	4	2.6666667	0.34	0.8499
noise#period#dial	11.333333	4	2.8333333	0.36	0.8357
Residual	127.11111	16	7.9444444		
Total	9924.8333	53	187.26101		

```
Between-subjects error term:  subject|noise
                     Levels:  6          (4 df)
      Lowest b.s.e. variable:  subject
      Covariance pooled over:  noise     (for repeated variables)
```

Repeated variable: period

```
                                Huynh-Feldt epsilon        =  1.0668
                                *Huynh-Feldt epsilon reset to 1.0000
                                Greenhouse-Geisser epsilon =  0.6476
                                Box's conservative epsilon =  0.5000
```

			Prob > F			
Source	df	F	Regular	H-F	G-G	Box
period	2	63.39	0.0000	0.0000	0.0003	0.0013
noise#period	2	5.67	0.0293	0.0293	0.0569	0.0759
period#subject|noise	8					

Repeated variable: dial

```
                                Huynh-Feldt epsilon        =  2.0788
                                *Huynh-Feldt epsilon reset to 1.0000
                                Greenhouse-Geisser epsilon =  0.9171
                                Box's conservative epsilon =  0.5000
```

			Prob > F			
Source	df	F	Regular	H-F	G-G	Box
dial	2	89.82	0.0000	0.0000	0.0000	0.0007
noise#dial	2	1.91	0.2102	0.2102	0.2152	0.2394
dial#subject|noise	8					

```
Repeated variables: period#dial
                                    Huynh-Feldt epsilon        =  1.3258
                                    *Huynh-Feldt epsilon reset to 1.0000
                                    Greenhouse-Geisser epsilon =  0.5134
                                    Box's conservative epsilon =  0.2500
                                                 ------------ Prob > F ------------
              Source |      df       F    Regular     H-F      G-G       Box
---------------------+--------------------------------------------------------------
         period#dial |       4    0.34     0.8499    0.8499   0.7295    0.5934
    noise#period#dial |       4    0.36     0.8357    0.8357   0.7156    0.5825
            Residual |      16
---------------------+--------------------------------------------------------------
```

For each repeated variable and for each combination of interactions of repeated variables, there are different ϵ correction values. The `anova` command produces tables for each applicable combination.

The two most significant factors in this model appear to be `dial` and `period`. The `noise` by `period` interaction may also be significant, depending on the correction factor you use. Below is a table of predictive margins for the accuracy score for dial, period, and noise by period.

```
. margins, within(dial)
Predictive margins                              Number of obs     =         54
Expression   : Linear prediction, predict()
within       : dial
Empty cells  : reweight
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dial |
           1 |   37.38889   .6643478    56.28   0.000     35.98053    38.79724
           2 |   42.22222   .6643478    63.55   0.000     40.81387    43.63058
           3 |   53.22222   .6643478    80.11   0.000     51.81387    54.63058
------------------------------------------------------------------------------

. margins, within(period)
Predictive margins                              Number of obs     =         54
Expression   : Linear prediction, predict()
within       : period
Empty cells  : reweight
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      period |
           1 |   54.33333   .6643478    81.78   0.000     52.92498    55.74169
           2 |      44.5    .6643478    66.98   0.000     43.09165    45.90835
           3 |         34   .6643478    51.18   0.000     32.59165    35.40835
------------------------------------------------------------------------------
```

```
. margins, within(noise period)
Predictive margins                              Number of obs     =         54
Expression   : Linear prediction, predict()
within       : noise period
Empty cells  : reweight
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
noise#period						
1 1	53.77778	.9395297	57.24	0.000	51.78606	55.76949
1 2	49.44444	.9395297	52.63	0.000	47.45273	51.43616
1 3	38.44444	.9395297	40.92	0.000	36.45273	40.43616
2 1	54.88889	.9395297	58.42	0.000	52.89717	56.8806
2 2	39.55556	.9395297	42.10	0.000	37.56384	41.54727
2 3	29.55556	.9395297	31.46	0.000	27.56384	31.54727

Dial shape 3 produces the highest score, and scores decrease over the periods.

◁

Example 17 had two repeated-measurement variables. Up to four repeated-measurement variables may be specified in the `anova` command.

Video examples

Analysis of covariance in Stata

Two-way ANOVA in Stata

Stored results

anova stores the following in e():

Scalars

e(N)	number of observations
e(mss)	model sum of squares
e(df_m)	model degrees of freedom
e(rss)	residual sum of squares
e(df_r)	residual degrees of freedom
e(r2)	R-squared
e(r2_a)	adjusted R-squared
e(F)	F statistic
e(rmse)	root mean squared error
e(ll)	log likelihood
e(ll_0)	log likelihood, constant-only model
e(ss_#)	sum of squares for term #
e(df_#)	numerator degrees of freedom for term #
e(ssdenom_#)	denominator sum of squares for term # (when using nonresidual error)
e(dfdenom_#)	denominator degrees of freedom for term # (when using nonresidual error)
e(F_#)	F statistic for term # (if computed)
e(N_bse)	number of levels of the between-subjects error term
e(df_bse)	degrees of freedom for the between-subjects error term
e(box#)	Box's conservative epsilon for a particular combination of repeated variables (repeated() only)
e(gg#)	Greenhouse–Geisser epsilon for a particular combination of repeated variables (repeated() only)
e(hf#)	Huynh–Feldt epsilon for a particular combination of repeated variables (repeated() only)
e(rank)	rank of e(V)

Macros

e(cmd)	anova
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(varnames)	names of the right-hand-side variables
e(term_#)	term #
e(errorterm_#)	error term for term # (when using nonresidual error)
e(sstype)	type of sum of squares; sequential or partial
e(repvars)	names of repeated variables (repeated() only)
e(repvar#)	names of repeated variables for a particular combination (repeated() only)
e(model)	ols
e(wtype)	weight type
e(wexp)	weight expression
e(properties)	b V
e(estat_cmd)	program used to implement estat
e(predict)	program used to implement predict
e(asbalanced)	factor variables fvset as asbalanced
e(asobserved)	factor variables fvset as asobserved

Matrices

e(b)	coefficient vector
e(V)	variance–covariance matrix of the estimators
e(Srep)	covariance matrix based on repeated measures (repeated() only)

Functions

e(sample)	marks estimation sample

References

Acock, A. C. 2018. *A Gentle Introduction to Stata*. 6th ed. College Station, TX: Stata Press.

Afifi, A. A., and S. P. Azen. 1979. *Statistical Analysis: A Computer Oriented Approach*. 2nd ed. New York: Academic Press.

Altman, D. G. 1991. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC.

Anderson, R. L. 1990. Gertrude Mary Cox 1900–1978. *Biographical Memoirs, National Academy of Sciences* 59: 116–132.

Box, G. E. P. 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* 25: 290–302.

Box, J. F. 1978. *R. A. Fisher: The Life of a Scientist*. New York: Wiley.

Chatfield, M. D., and A. P. Mander. 2009. The Skillings–Mack test (Friedman test when there are missing data). *Stata Journal* 9: 299–305.

Cobb, G. W. 1998. *Introduction to Design and Analysis of Experiments*. New York: Springer.

Edwards, A. L. 1985. *Multiple Regression and the Analysis of Variance and Covariance*. 2nd ed. New York: Freeman.

Fisher, R. A. 1925. *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

——. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

——. 1990. *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford: Oxford University Press.

Geisser, S., and S. W. Greenhouse. 1958. An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics* 29: 885–891.

Gleason, J. R. 1999. sg103: Within subjects (repeated measures) ANOVA, including between subjects factors. *Stata Technical Bulletin* 47: 40–45. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 236–243. College Station, TX: Stata Press.

——. 2000. sg132: Analysis of variance from summary statistics. *Stata Technical Bulletin* 54: 42–46. Reprinted in *Stata Technical Bulletin Reprints*, vol. 9, pp. 328–332. College Station, TX: Stata Press.

Hall, N. S. 2010. Ronald Fisher and Gertrude Cox: Two statistical pioneers sometimes cooperate and sometimes collide. *American Statistician* 64: 212–220.

Higgins, J. E., and G. G. Koch. 1977. Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review* 45: 51–62.

Huber, C. 2013. Measures of effect size in Stata 13. *The Stata Blog: Not Elsewhere Classified*. http://blog.stata.com/2013/09/05/measures-of-effect-size-in-stata-13/.

Huynh, H. 1978. Some approximate tests for repeated measurement designs. *Psychometrika* 43: 161–175.

Huynh, H., and L. S. Feldt. 1976. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics* 1: 69–82.

Kennedy, W. J., Jr., and J. E. Gentle. 1980. *Statistical Computing*. New York: Dekker.

Lalanne, C., and M. Mesbah. 2016. *Biostatistics and Computer-based Analysis of Health Data Using Stata*. London: ISTE Press.

Marchenko, Y. V. 2006. Estimating variance components in Stata. *Stata Journal* 6: 1–21.

Mehmetoglu, M., and T. G. Jakobsen. 2017. *Applied Statistics Using Stata: A Guide for the Social Sciences*. Thousand Oaks, CA: Sage.

Milliken, G. A., and D. E. Johnson. 2009. *Analysis of Messy Data, Volume 1: Designed Experiments*. 2nd ed. Boca Raton, FL: CRC Press.

Mitchell, M. N. 2012. *Interpreting and Visualizing Regression Models Using Stata*. College Station, TX: Stata Press.

——. 2015. *Stata for the Behavioral Sciences*. College Station, TX: Stata Press.

Mooi, E., M. Sarstedt, and I. Mooi-Reci. 2018. *Market Research: The Process, Data, and Methods Using Stata*. Singapore: Springer.

Scheffé, H. 1959. *The Analysis of Variance*. New York: Wiley.

Snedecor, G. W., and W. G. Cochran. 1989. *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press.

van Belle, G., L. D. Fisher, P. J. Heagerty, and T. S. Lumley. 2004. *Biostatistics: A Methodology for the Health Sciences*. 2nd ed. New York: Wiley.

Weinberg, S. L., and S. K. Abramowitz. 2016. *Statistics Using Stata: An Integrative Approach*. New York: Cambridge University Press.

Winer, B. J., D. R. Brown, and K. M. Michels. 1991. *Statistical Principles in Experimental Design*. 3rd ed. New York: McGraw–Hill.

Also see

[R] **anova postestimation** — Postestimation tools for anova

[R] **contrast** — Contrasts and linear hypothesis tests after estimation

[R] **icc** — Intraclass correlation coefficients

[R] **loneway** — Large one-way ANOVA, random effects, and reliability

[R] **oneway** — One-way analysis of variance

[R] **regress** — Linear regression

[MV] **manova** — Multivariate analysis of variance and covariance

[PSS] **power oneway** — Power analysis for one-way analysis of variance

[PSS] **power repeated** — Power analysis for repeated-measures analysis of variance

[PSS] **power twoway** — Power analysis for two-way analysis of variance

Stata Structural Equation Modeling Reference Manual

[U] **20 Estimation and postestimation commands**