The Stata Journal (2006)
6, Number 1, pp. 1–21

Estimating variance components in Stata

Yulia Marchenko
StataCorp
College Station, TX
ymarchenko@stata.com

Abstract. This article gives a brief overview of the popular methods for estimating variance components in linear models and describes several ways to obtain such estimates in Stata for various experimental designs. The article's emphasis is on using `xtmixed` to estimate variance components. Prior to Stata 9, `loneway` could be used to estimate variance components for one-way random-effects models. For other experimental designs, variance components could be computed manually using saved results after `anova`. The latter approach is viable but requires tedious computations for complicated experimental designs. Instead, as of Stata 9, variance components are easily obtained by using `xtmixed`.

Keywords: st0095, variance components, experimental design, ANOVA, REML, ML, multilevel, random coefficients, mixed models

1 Introduction

Various methods exist for estimating variance components. Among them are analysis of variance (ANOVA), maximum likelihood (ML), restricted maximum likelihood (REML), minimum norm, and Bayes. For a history of methods for estimating variance components, see Searle, Casella, and McCulloch (1992). This article concentrates on how to obtain variance components in Stata using the ANOVA, REML, and ML methods.

The general method for estimating variance components by equating ANOVA mean squares to their expected values, known as the ANOVA method of estimation, is due to Tippett (1931). Several adaptations of the ANOVA method for unbalanced data were proposed by Henderson (1953). The algorithms for computing ANOVA estimates of variance components for both balanced and unbalanced data are discussed in Searle, Casella, and McCulloch (1992).

ANOVA estimation of variance components involves solving a system of linear equations, with the structure of the system dependent on the specific experimental design. As such, a general program to compute ANOVA-type estimates is, at best, a difficult concept. I do, however, demonstrate this method in section 2 for one specific design.

Serious weaknesses of ANOVA estimators—for example, possibly negative estimates of variance components, nonexistence of uniformly best estimators, and lack of uniqueness in the case of unbalanced data—have led to the investigation of alternative methods of variance components estimation. Two alternatives are ML (Hartley and Rao 1967) and REML (Thompson 1992). These methods are based on maximizing the likelihood function corresponding to the statistical model that underlies the experimental design;

they require a distributional assumption on the response, i.e., normality. The REML method is based on maximizing the portion of the likelihood that is invariant to the fixed effects. The REML and ML estimates are guaranteed to be nonnegative. The difference between ML and REML estimators is that the latter takes into account the implicit degrees of freedom associated with the fixed effects. For balanced designs, ANOVA and REML estimators are identical. For unbalanced designs, all three estimators generally differ. Because of their simplicity relative to ANOVA methods, ML and REML are the preferred methods of estimation for unbalanced data.

As of Stata 9, you can obtain ML and REML estimates of variance components by using `xtmixed`. The key, however, lies in expressing the various experimental designs as multilevel mixed-effects models, i.e., in the language used by `xtmixed`.

Section 2 describes the ANOVA method for estimating variance components and demonstrates how ANOVA-type estimates can be obtained using Stata. Section 3 discusses `xtmixed` as a tool for variance-components estimation. Section 4 provides examples of how to get variance components estimates in Stata for several experimental designs.

2 ANOVA-type estimation of variance components

We demonstrate two methods of computing ANOVA-type estimates of variance components manually after `anova` for a random two-way full factorial experimental design.

ANOVA-type estimates of variance components can be obtained by solving the linear-equation system obtained from equating the expected mean squares to their sample estimates, which are labeled in `anova` output as "mean squares". We can define $\mathbf{b}$ to be the column vector of mean squares and matrix $\mathbf{C}$ to be the matrix of coefficients that links expected mean squares to observed mean squares. The structure of matrix $\mathbf{C}$ depends on a particular experimental design. Let $\mathbf{v}$ be the column vector of unknown variance components. Then $\mathbf{v}$ is a solution to

$$\mathbf{C}\mathbf{v} = \mathbf{b}$$

As such, one method for estimating variance components is to use the Stata matrix commands to construct the matrices $\mathbf{b}$ and $\mathbf{C}$ and to compute components of $\mathbf{v}$ as

$$\mathbf{v} = \mathbf{C}^{-1}\mathbf{b} \tag{1}$$

You can also directly use formulas readily available for common experimental designs to compute variance components; see, for example, Kuehl (2000); Winer, Brown, and Michels (1991); and Searle, Casella, and McCulloch (1992). However, such formulas are merely a more direct representation of (1).

2.1 ANOVA-type estimates for random-effects two-way full factorial design

As an example to show how to compute estimates of variance components after `anova` by using the two methods described above for a random two-factor full factorial design, we use the data from example 7.1 in Kuehl (2000). The measurements on triglyceride levels (milligrams per deciliter) in the serum samples were obtained from a randomly selected sample of machines to evaluate machine performance. The research problem is to estimate the variability of measurements among machines operated over several days. Four machines ($b = 4$) were selected for the study, with two measurements ($r = 2$) obtained from each machine for each of the 4 days ($a = 4$). The sources of variation are variability among machines, σ^2_m; variability among days, σ^2_d; variability associated with interaction between days and machines, σ^2_{dm}; and error variability, σ^2_e.

We fit this design using `anova` and obtain variance components directly by using published formulas and by solving the system of linear equations. Here `trigly` is the dependent variable; `day` and `machine` define random factors.

```
. use trigly1
(Kuehl, example 7.1 (trigly data))

. anova trigly day machine day*machine

                         Number of obs =      32     R-squared     =  0.9294
                         Root MSE      = 4.23028     Adj R-squared =  0.8632

                  Source |  Partial SS    df       MS           F     Prob > F
             ------------+----------------------------------------------------
                   Model |  3767.77723    15  251.185149      14.04     0.0000
                         |
                     day |  1334.46338     3  444.821125      24.86     0.0000
                 machine |  1647.27875     3  549.092916      30.68     0.0000
             day*machine |  786.035104     9  87.3372338       4.88     0.0029
                         |
                Residual |  286.324902    16  17.8953064
             ------------+----------------------------------------------------
                   Total |  4054.10213    31  130.777488
```

The first method is to compute estimates of variance components for terms `day`, `machine`, and `day*machine` directly using the formulas

$$\widehat{\sigma}^2_r = \mathrm{MS}(\texttt{Residual})$$

$$\widehat{\sigma}^2_{dm} = \frac{\mathrm{MS}(\texttt{day*machine}) - \mathrm{MS}(\texttt{Residual})}{r}$$

$$\widehat{\sigma}^2_m = \frac{\mathrm{MS}(\texttt{machine}) - \mathrm{MS}(\texttt{day*machine})}{ra}$$

$$\widehat{\sigma}^2_d = \frac{\mathrm{MS}(\texttt{day}) - \mathrm{MS}(\texttt{day*machine})}{rb}$$

and the values of sum of squares saved after `anova` as shown below. Since sums of squares are what are saved in `e()` after `anova`, they must be converted to mean squares by dividing by the appropriate degrees of freedom.

```
. local a = 4
. local b = 4
. local r = 2
. local resid = e(rss)/e(df_r)
. local dayXmachine = (e(ss_3)/e(df_3) - `resid')/`r'
. local mach  = (e(ss_2)/e(df_2) - e(ss_3)/e(df_3))/(`r'*`a')
. local day  = (e(ss_1)/e(df_1) - e(ss_3)/e(df_3))/(`r'*`b')
. display as txt "Variance components:"
Variance components:
. display as txt "Var(day) = " as res `day'
Var(day) = 44.685486
. display as txt "Var(machine) = " as res `mach'
Var(machine) = 57.71946
. display as txt "Var(dayXmachine) = " as res `dayXmachine'
Var(dayXmachine) = 34.720964
. display as txt "Var(residual) = " as res `resid'
Var(residual) = 17.895306
```

In matrix notation, we have the following system of linear equations to estimate variance components corresponding to this experimental design:

$$\begin{pmatrix} rb & 0 & r & 1 \\ 0 & ra & r & 1 \\ 0 & 0 & r & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_d^2 \\ \sigma_m^2 \\ \sigma_{dm}^2 \\ \sigma_r^2 \end{pmatrix} = \begin{pmatrix} \text{MS}(\texttt{day}) \\ \text{MS}(\texttt{machine}) \\ \text{MS}(\texttt{day*machine}) \\ \text{MS}(\texttt{Residual}) \end{pmatrix}$$

We can obtain the solution to this system of linear equations in Stata as follows:

```
. mat b = (e(ss_1)/e(df_1) \ e(ss_2)/e(df_2) \ e(ss_3)/e(df_3) \e(rss)/e(df_r))
. mat C = (`r'*`b',0,`r',1\0,`r'*`a',`r',1\0,0,`r',1\0,0,0,1)
. mat v = inv(C)*b
. mat v = v'
. matrix rownames v = Var
. matrix colnames v = day machine dayXmachine residual
. display as txt "Variance components:"
Variance components:
. mat list v
v[1,4]
             day      machine  dayXmachine     residual
Var    44.685486     57.71946    34.720964    17.895306
```

The structure of the matrix C can become complicated and can involve tedious computation of its entries for unbalanced and complicated experimental designs,[1] which makes manually computing variance components more difficult.

1. Expressions for the variance components for some such designs can be found in Searle, Casella, and McCulloch (1992) and Winer, Brown, and Michels (1991).

As of Stata 9, variance components for such designs can be easily estimated with `xtmixed`. We demonstrate this process for this particular design in section 4.2.

3 xtmixed as a tool for variance component estimation

REML and ML estimates of variance components can be obtained in Stata by using `xtmixed` for both balanced and unbalanced designs. By default, `xtmixed` produces REML estimates. You can obtain ML estimates by using option `mle`.

I draw the attention of ANOVA-oriented Stata users to `xtmixed` as a tool for variance-components estimation for random-effects experimental designs. `xtmixed` is designed primarily for fitting random coefficients and multilevel models. However, statistical models underlying random-effects experimental designs can be viewed as particular types of multilevel models. For example, a one-way random-effects experimental design corresponds to a random-intercept model; the experimental design with two nested random factors can be treated as the two-level random-intercept model.

The difference between the ANOVA and multilevel representations of the models is in the organization of the data. In multilevel models, the data are viewed as a series of independent panels where each panel contains a vector of responses, with the specified covariance structure, $\mathbf{\Sigma}$, of random effects, $\mathbf{u}$, where $\mathbf{u}$ is independently observed within each panel. On the other hand, an ANOVA specification considers all n observations at once, with corresponding covariance matrix[2] $\mathbf{G} = I_M \otimes \mathbf{\Sigma}$ of random effects, where M defines the number of panels (for the specification of the models corresponding to the two representations discussed above, refer to [XT] **xtmixed**). An ANOVA representation of the model corresponds to treating all data as one big panel with a certain block-diagonal covariance structure.

Since variance components, along with error variance σ_e^2, are characterized by elements of the matrix $\mathbf{G}$ and therefore by elements of matrix $\mathbf{\Sigma}$, they are the same for both ANOVA and multilevel model formulations. The latter, however, is more computationally efficient because of the lower dimension of the design matrix for random effects, $\mathbf{u}$.

The design-matrix–based approach, or as we call it, the brute-force way of fitting random-effects designs with `xtmixed`, is to construct the design matrix for random effects in a straightforward way by specifying indicator variables corresponding to the levels of all random effects. In multilevel language, this approach means considering all data as one big group, treating random factors as being nested within this group, and treating levels of random factors as random coefficients on indicator variables for these random factors. The random coefficients are assumed (a) to have equal variances within a random effect, (b) to be uncorrelated among each other, and (c) to be uncorrelated with the random coefficients for other random effects. To accommodate the design-matrix–based approach, `xtmixed` supports the special group identifier `_all` and the factor notation `R.`*varname* (see [XT] **xtmixed**). The syntax for `xtmixed` corresponding to the brute-force way of fitting random-effects designs is

2. $\otimes$ denotes the Kronecker product of two matrices.

```
. xtmixed depvar fe_equation [|| _all: R.re_varname1] [|| _all: R.re_varname2 ...]
```

where *fe_equation* includes fixed effects defining a regression function; `_all` corresponds to the ID variable identifying all the observations as one big panel; and `R.`*re_varname1*, `R.`*re_varname2*, and so on define random-effects variables *re_varname1*, *re_varname2* as factor variables. When `R.`-notation is used, by default, the identity covariance structure is specified for the random effects. This condition fulfills requirements (a) and (b). Also, since random factors from different random equations are independent, assumption (c) is achieved by listing each random factor in a separate random equation. The syntax above corresponds to the ANOVA formulation of the model.

However, such a direct approach can be computationally burdensome. That is, since the `R.`-notation defines each of the levels of the random factor as a separate parameter in the vector of random effects, the column dimension of the design matrix for the random effects is increased. For example, in the specification above, the column dimension of the design matrix for random effects is equal to the total number of levels of each random effect. When the number of levels is very large, the consequences may be an increase in the computation time and a failure to accommodate enough memory required for fitting complicated experimental designs.

In such situations, formulating the model as a multilevel model is advantageous since it results in a significant reduction of the dimensionality of the random effects. For example, random-effects–nested designs with all nested factors being random correspond to the random-intercept multilevel model with levels defined by these random factors:

```
. xtmixed depvar fe_equation [|| re_varname1:] [|| re_varname2: ...]
```

where *re_varname1* defines first-level groups; *re_varname2*, being nested within *re_varname1*, defines second-level groups; and so on. The column dimension of the design matrix for random effects in this case is equal to the number of random factors. For the random-effects designs with crossed factors, we cannot avoid using `R.`-notation. However, as we show later in our examples, there are more effective ways of fitting such designs than the brute-force way.

In what follows, I demonstrate examples of using `xtmixed` effectively to get estimates of variance components for different experimental designs. A detailed discussion of using `xtmixed` for random-effects models and ways to fit them more effectively is given in Rabe-Hesketh and Skrondal (2005). A general description of multilevel models can be found in Goldstein (2003). See also [XT] **xtmixed** for a general description of that command.

4 Examples

I demonstrate how to obtain estimates of variance components for several experimental designs using `xtmixed`. I provide both the brute-force way of using `xtmixed` with the direct translation of an ANOVA model and the more efficient way of obtaining the same results with `xtmixed` using a multilevel model specification.

4.1 Random effects for one-factor experimental design

Here I demonstrate how to obtain variance-components estimates for a single random-factor experimental design using both `loneway` and `xtmixed`. The data for this example are taken from example 5.1 in Kuehl (2000). Tensile-strength measurements of the alloy are obtained on a random sample of 10 ($r = 10$) bars from each of the three castings ($t = 3$). The research objective of the experiment is to study the variability among the bars, σ_e^2, taking into account possible variability due to different castings, σ_t^2.

We first estimate the variance components σ_t^2 and σ_e^2 by using `loneway`. The variable `temp` is the dependent variable and `casting` defines a random factor. Error variability defines the source of the variability among bars.

Using `loneway`, we type

```
. use alloy
(Kuehl, example 5.1 (alloy data))

. loneway temp casting
                   One-way Analysis of Variance for temp:

                                           Number of obs =        30
                                               R-squared =    0.4849

    Source                SS         df      MS            F     Prob > F
-------------------------------------------------------------------------

Between casting       147.88464      2     73.94232      12.71     0.0001
Within casting        157.10202     27    5.8185932
-------------------------------------------------------------------------

Total                 304.98666     29    10.516781

         Intraclass       Asy.
         correlation      S.E.       [95% Conf. Interval]
         ------------------------------------------------
            0.53934     0.27948      0.00000     1.08712

         Estimated SD of casting effect          2.610052
         Estimated SD within casting             2.412176
         Est. reliability of a casting mean       0.92131
              (evaluated at n=10.00)
```

The estimated variance components can be obtained as the square of the corresponding estimated standard deviations. The estimate of variability among bars is $\widehat{\sigma}_e^2 = (2.412)^2 = 5.82$, and the estimate of variability among castings is $\widehat{\sigma}_t^2 = (2.610)^2 = 6.81$.

Now we use `xtmixed` to estimate these same variance components. By default, `xtmixed` reports these as standard deviations, but we can specify option `variance` to get estimates of variances.

The direct translation of the ANOVA model corresponds to specifying indicator variables for each level of the random effect, `casting` in our example, which corresponds to the following syntax for `xtmixed`:

```
. xtmixed temp || _all: R.casting, variance nolog

Mixed-effects REML regression                   Number of obs      =        30
Group variable: _all                            Number of groups   =         1

                                                Obs per group: min =        30
                                                               avg =      30.0
                                                               max =        30

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -70.927391          Prob > chi2        =         .

------------------------------------------------------------------------------
        temp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   90.86667   1.569951    57.88   0.000     87.78962    93.94371
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
_all: Identity               |
              var(R.casting) |   6.812376   7.395932      .8113012    57.20251
-----------------------------+------------------------------------------------
               var(Residual) |   5.818593    1.58362       3.41311    9.919407
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    12.08 Prob >= chibar2 = 0.0003
```

The column dimension of the design matrix for random effects corresponding to this syntax is equal to 3, the number of levels of `casting`. The more efficient way to do this is to fit this model as a one-level random-intercept model with `casting` as a group variable with a random intercept for each group. This method reduces the column dimension of the design matrix to 1:

```
. xtmixed temp || casting:, variance nolog

Mixed-effects REML regression                   Number of obs      =        30
Group variable: casting                         Number of groups   =         3

                                                Obs per group: min =        10
                                                               avg =      10.0
                                                               max =        10

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -70.927391          Prob > chi2        =         .

------------------------------------------------------------------------------
        temp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   90.86667   1.569951    57.88   0.000     87.78962    93.94371
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
casting: Identity            |
                   var(_cons) |   6.812376   7.395932      .8113011    57.20252
-----------------------------+------------------------------------------------
               var(Residual) |   5.818593    1.58362       3.41311    9.919407
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =    12.08 Prob >= chibar2 = 0.0003
```

The two estimations produce identical results; however, the advantage of the second specification would become apparent when the number of levels of the random effect is very large.

4.2 Random effects for two-way full factorial experimental design

Let us now go back to the example from section 2.1.

In accordance with the factor notation of the corresponding ANOVA model, we can use the following syntax of `xtmixed`. There is no automatic way to specify an interaction variable within `xtmixed`, but we can create the appropriate group variable manually by using `egen`:

```
. use trigly1
(Kuehl, example 7.1 (trigly data))
. egen dayXmachine=group(machine day)
. xtmixed trigly || _all: R.day || _all: R.machine || _all: R.dayXmachine,
> variance nolog
Mixed-effects REML regression                   Number of obs      =        32
Group variable: _all                            Number of groups   =         1
                                                Obs per group: min =        32
                                                               avg =      32.0
                                                               max =        32
                                                Wald chi2(0)       =         .
Log restricted-likelihood = -107.51918          Prob > chi2        =         .

      trigly |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   141.1844   5.322644    26.53   0.000     130.7522    151.6166

  Random-effects Parameters |   Estimate   Std. Err.     [95% Conf. Interval]
----------------------------+------------------------------------------------
_all: Identity              |
                 var(R.day) |   44.68551   45.69017      6.023203     331.517
----------------------------+------------------------------------------------
_all: Identity              |
             var(R.machine) |    57.7195   56.27743      8.538616    390.1734
----------------------------+------------------------------------------------
_all: Identity              |
            var(R.dayXma~e) |   34.72102   20.82727      10.71526    112.5077
----------------------------+------------------------------------------------
              var(Residual) |   17.89529   6.326937      8.949396    35.78358

LR test vs. linear regression:       chi2(3) =    27.48   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference
```

The estimated variance components are identical to those derived using ANOVA methods in section 2.1.

The more efficient way to fit this model is as a three-level model with crossed terms. Here the factors `day` and `machine` are crossed. As described in Goldstein (2003) and

Rabe-Hesketh and Skrondal (2005), we can specify the model corresponding to this two-way random full factorial experimental design as follows.

1. Treat both factors to be nested within the entire dataset.
2. Choose one of the factors, usually the one with the largest number of levels, to define a random intercept at the second level.
3. Create a set of indicator explanatory variables, one for each category, for the other factor at the first level with random intercepts uncorrelated and with variances constrained to be equal. This step can be done automatically by using the `R.`*varname* notation.
4. Use an additional nesting level to estimate the variance component for the interaction term.

```
. xtmixed trigly || _all: R.day || machine: || dayXmachine:, variance nolog
Mixed-effects REML regression                   Number of obs      =        32

-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
           _all |        1         32       32.0         32
        machine |        4          8        8.0          8
    dayXmachine |       16          2        2.0          2
-----------------------------------------------------------

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -107.51918          Prob > chi2        =         .

------------------------------------------------------------------------------
      trigly |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   141.1844   5.322644    26.53   0.000     130.7522    151.6166
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
_all: Identity               |
                  var(R.day) |   44.68551   45.69017      6.023203     331.517
-----------------------------+------------------------------------------------
machine: Identity            |
                  var(_cons) |    57.7195   56.27743      8.538615    390.1734
-----------------------------+------------------------------------------------
dayXmachine: Identity        |
                  var(_cons) |   34.72102   20.82727      10.71526    112.5077
-----------------------------+------------------------------------------------
               var(Residual) |   17.89529   6.326937      8.949396    35.78358
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =    27.48   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference
```

Since the interaction can also be viewed as nesting one factor within another, you can also fit the above model by using

```
. xtmixed trigly || _all: R.day || machine: || day:, variance nolog
Mixed-effects REML regression                   Number of obs      =        32

                    No. of       Observations per Group
 Group Variable     Groups    Minimum    Average    Maximum

           _all          1         32       32.0         32
        machine          4          8        8.0          8
            day         16          2        2.0          2

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -107.51918          Prob > chi2        =         .

      trigly        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       _cons     141.1844   5.322644    26.53   0.000     130.7522    151.6166


  Random-effects Parameters      Estimate   Std. Err.     [95% Conf. Interval]

_all: Identity
                  var(R.day)     44.68551   45.69017      6.023203     331.517

machine: Identity
                  var(_cons)      57.7195   56.27743      8.538615    390.1734

day: Identity
                  var(_cons)     34.72102   20.82727      10.71526    112.5077

               var(Residual)     17.89529   6.326937      8.949396    35.78358

LR test vs. linear regression:       chi2(3) =    27.48   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference
```

which does not require creating a separate interaction variable. All these estimations using `xtmixed` are identical, yet the final way is the most efficient because the design matrix for random effects is of lower dimension and we avoid creating an interaction variable. The first estimation requires a design matrix for random effects with column dimension equal to $4 + 4 + 4 \times 4 = 24$, whereas the other two need only $4 + 1 + 1 = 6$ random-effects parameters. In the second estimation we also need to create an interaction variable.

4.3 Random effects for mixed experimental design with crossed factors

Here I give an example of how to use `xtmixed` to estimate variance components for the two-way full factorial mixed design. The data are obtained from example 7.2 in Kuehl (2000). Two measurements of triglyceride levels (milligrams per deciliter) ($r = 2$) are obtained for each of the two methods ($a = 2$) on each of the 4 days ($b = 4$). Here `method` is a fixed factor, `day` is a random factor, and interaction between method and day, `method*day`, is also a random factor.

The ANOVA table obtained for this experiment is as follows:

```
. use trigly2
(Kuehl, example 7.2 (trigly data))

. anova trigly method day method*day

                           Number of obs =      16     R-squared     =  0.8913
                           Root MSE      = 3.79704     Adj R-squared =  0.7962

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  945.697524     7  135.099646       9.37     0.0026
                         |
                  method |  329.422694     1  329.422694      22.85     0.0014
                     day |  431.442437     3  143.814146       9.97     0.0044
              method*day |  184.832393     3  61.6107977       4.27     0.0446
                         |
                Residual |  115.340023     8  14.4175029
              -----------+----------------------------------------------------
                   Total |  1061.03755    15  70.7358365
```

Using the values of mean squares given in Kuehl (2000), you can calculate the variance components to be $\widehat{\sigma}_e^2 = 14$ for the error (`Residual`), $\widehat{\sigma}_{dm}^2 = (62 - 14)/2 = 24$ for the interaction (`method*day`), and $\widehat{\sigma}_d^2 = (144 - 62)/(2 \times 2) = 20.5$ for the day (`day`) terms. Below is an example of using `xtmixed` efficiently to estimate variance components for this design. Here we again define an interaction through a nesting of factors.

```
. xi: xtmixed trigly i.method || day: || method:, variance nolog
i.method          _Imethod_1-2        (naturally coded; _Imethod_1 omitted)

Mixed-effects REML regression                   Number of obs      =        16

-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
            day |        4          4        4.0          4
         method |        8          2        2.0          2
-----------------------------------------------------------

                                                Wald chi2(1)       =      5.35
Log restricted-likelihood = -46.252391          Prob > chi2        =    0.0208

------------------------------------------------------------------------------
      trigly |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imethod_2 |  -9.075003   3.924627    -2.31   0.021    -16.76713   -1.382875
       _cons |        147   3.583163    41.03   0.000     139.9771    154.0229
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
day: Identity                |
                  var(_cons) |   20.55083    31.9364      .9773392     432.129
-----------------------------+------------------------------------------------
method: Identity             |
                  var(_cons) |   23.59664   25.40945      2.859271    194.7355
-----------------------------+------------------------------------------------
               var(Residual) |   14.41751   7.208753      5.411147    38.41412
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =     6.77   Prob > chi2 = 0.0339

Note: LR test is conservative and provided only for reference
```

The same results can be obtained using the brute-force specification of `xtmixed`:

```
. egen dayXmethod = group(day method)
. xi: xtmixed trigly i.method || _all: R.day || _all: R.dayXmethod, variance nolog
```

4.4 Nested-factor experimental design

The data come from example 7.3 in Kuehl (2000). Glucose measurements (milligrams per deciliter) were collected to study the performance of serum assays critical for the correct medical diagnoses. The important sources of variation on the assays are days on which the assays are conducted, σ_a^2; the replicate runs within days, $\sigma_{b(a)}^2$; and the replicate serum sample preparations within run, $\sigma_{c(b)}^2$. There are three ($c = 3$) replications of glucose standards prepared for each of two ($b = 2$) runs on each of 3 ($a = 3$) days. This is an example of the nested experimental design with three random nested factors: day (`day`), run|day (`run|day`), and rep|run (`Residual`).

First, we use `anova` to produce a table corresponding to this design:

```
. use glucose
(Kuehl, example 7.3 (glucose data))
. anova glucose day / run|day /

                           Number of obs =      18     R-squared     =  0.6864
                           Root MSE      = 1.07083     Adj R-squared =  0.5558

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  30.1200012     5  6.02400023       5.25     0.0087
                         |
                     day |  13.7633271     2  6.88166354       1.26     0.4002
                 run|day |  16.3566741     3   5.4522247
              -----------+----------------------------------------------------
                 run|day |  16.3566741     3   5.4522247       4.75     0.0208
                         |
                Residual |  13.7600005    12   1.1466667
              -----------+----------------------------------------------------
                   Total |  43.8800016    17  2.58117657
```

We demonstrate the brute-force way of fitting `xtmixed` to obtain variance components for this design. Since runs are nested within days, we cannot estimate variability due to runs only and, therefore, we cannot use the `R.`*run* notation to define the random effects for estimating $\sigma_{c(b)}^2$. Instead, we must create an interaction between run and day and use it with `R.`-notation:

(*Continued on next page*)

```
. egen dayXrun = group(day run)
. xtmixed glucose || _all: R.day || _all: R.dayXrun, variance nolog
Mixed-effects REML regression                   Number of obs      =        18
Group variable: _all                            Number of groups   =         1
                                                Obs per group: min =        18
                                                               avg =      18.0
                                                               max =        18
                                                Wald chi2(0)       =         .
Log restricted-likelihood = -30.861192          Prob > chi2        =         .

     glucose        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       _cons     42.76667   .6183155    69.17   0.000     41.55479    43.97854


  Random-effects Parameters       Estimate   Std. Err.     [95% Conf. Interval]

_all: Identity
                  var(R.day)      .2382376   1.366002      3.14e-06     18097.3

_all: Identity
              var(R.dayXrun)      1.435187   1.492089      .1870504     11.0118

               var(Residual)      1.146667   .4681248      .5151523    2.552342

LR test vs. linear regression:       chi2(2) =     5.53   Prob > chi2 = 0.0629
Note: LR test is conservative and provided only for reference
```

Now we use `xtmixed` more efficiently by fitting the model for a nested random-effects design as a two-level random-intercept model:

```
. xtmixed glucose || day: || run:, variance nolog
Mixed-effects REML regression                   Number of obs      =        18

                    No. of       Observations per Group
 Group Variable     Groups    Minimum    Average    Maximum

            day          3          6        6.0          6
            run          6          3        3.0          3

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -30.861192          Prob > chi2        =         .

     glucose        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       _cons     42.76667   .6183155    69.17   0.000     41.55479    43.97854
```

```
Random-effects Parameters    Estimate   Std. Err.     [95% Conf. Interval]

day: Identity
                var(_cons)   .2382376   1.366002      3.14e-06     18097.3

run: Identity
                var(_cons)   1.435187   1.492089      .1870504     11.0118

             var(Residual)   1.146667   .4681248      .5151523    2.552342

LR test vs. linear regression:      chi2(2) =     5.53   Prob > chi2 = 0.0629
Note: LR test is conservative and provided only for reference
```

You can obtain variance components for this design by specifying only one random-effects equation. This goal can be achieved by noting that the covariance matrix of the data is block-diagonal with exchangeable matrices on the diagonal blocks. We can thus fit the same model as follows:

```
. xtmixed glucose || day: R.run, cov(exchangeable) variance nolog
Mixed-effects REML regression                   Number of obs      =        18
Group variable: day                             Number of groups   =         3

                                                Obs per group: min =         6
                                                               avg =       6.0
                                                               max =         6

                                                Wald chi2(0)       =         .
Log restricted-likelihood = -30.861192          Prob > chi2        =         .

     glucose       Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

       _cons    42.76667    .618316    69.17   0.000     41.55479    43.97854


Random-effects Parameters    Estimate   Std. Err.     [95% Conf. Interval]

day: Exchangeable
                var(R.run)   1.673426   1.374891       .334393     8.37444
                cov(R.run)   .2382399   1.366007     -2.439084    2.915564

             var(Residual)   1.146667   .4681248      .5151523    2.552342

LR test vs. linear regression:      chi2(2) =     5.53   Prob > chi2 = 0.0629
Note: LR test is conservative and provided only for reference
```

The corresponding variance components are $\widehat{\sigma}^2_a = \texttt{cov(R.run)} = .238$, $\widehat{\sigma}^2_{b(a)} = \texttt{var(R.run)} - \texttt{cov(R.run)} = 1.435$, and $\widehat{\sigma}^2_{c(b)} = 1.147$, which agree with previous results. For a detailed explanation, see example 7 in [XT] **xtmixed**.

Being able to estimate variance components for two nested factors, one nested within another, in one equation is handy for fitting random-effects designs with nested and crossed factors, as I demonstrate in subsection 4.6.

4.5 Nested-factor mixed experimental design

Now we fit a mixed experimental design with a nested factor, assuming that `day` is a fixed factor in the example described in section 4.4. The direct ANOVA formulation of the model requires that we specify random coefficients on indicator variables for `run` within each level of `day`:

```
. use glucose
(Kuehl, example 7.3 (glucose data))

. xi: xtmixed glucose i.day || day: R.run, variance nolog
i.day             _Iday_1-3          (naturally coded; _Iday_1 omitted)

Mixed-effects REML regression                    Number of obs      =        18
Group variable: day                              Number of groups   =         3

                                                 Obs per group: min =         6
                                                                avg =       6.0
                                                                max =         6

                                                 Wald chi2(2)       =      2.52
Log restricted-likelihood = -27.336908          Prob > chi2        =    0.2830

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iday_2 |   1.633333   1.348113     1.21   0.226     -1.00892    4.275585
     _Iday_3 |  -.3833338   1.348113    -0.28   0.776    -3.025587    2.258919
       _cons |      42.35   .9532598    44.43   0.000     40.48165    44.21836
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
day: Identity                |
                  var(R.run) |   1.435186    1.49209      .1870499    11.01182
-----------------------------+------------------------------------------------
               var(Residual) |   1.146667   .4681247      .5151523    2.552341
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =     3.73 Prob >= chibar2 = 0.0268
```

Interchanging the roles of `run` and `day` does not affect estimation results, so it is more efficient to specify the factor with fewer levels with the `R.`-notation. For example, if `day` had fewer levels than `run`, the following syntax would result in a smaller column dimension of the design matrix for random effects:

```
. xi: xtmixed glucose i.day || run: R.day, variance nolog
```

A more efficient way to obtain the results above is to express this design as a one-level random-intercept model with the level defined by the interaction between `day` and `run`:

```
. egen dayXrun = group(day run)

. xi: xtmixed glucose i.day || dayXrun:, variance nolog
i.day             _Iday_1-3          (naturally coded; _Iday_1 omitted)

Mixed-effects REML regression                   Number of obs      =        18
Group variable: dayXrun                         Number of groups   =         6

                                                Obs per group: min =         3
                                                               avg =       3.0
                                                               max =         3

                                                Wald chi2(2)       =      2.52
Log restricted-likelihood = -27.336908          Prob > chi2        =    0.2830

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iday_2 |   1.633333   1.348113     1.21   0.226     -1.00892    4.275585
     _Iday_3 |  -.3833338   1.348113    -0.28   0.776    -3.025587    2.258919
       _cons |      42.35   .9532598    44.43   0.000     40.48165    44.21836
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
dayXrun: Identity            |
                  var(_cons) |   1.435186    1.49209        .18705    11.01182
-----------------------------+------------------------------------------------
               var(Residual) |   1.146667   .4681247      .5151523    2.552341
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =     3.73 Prob >= chibar2 = 0.0268
```

The alternative specification of the model above comes in handy when, for example, we want to include a random coefficient for some covariate, x, that is measured within the levels of the random factor. For the above example, if some covariate, x, is measured within levels of factor run, the syntax below can be used to fit the model:

```
. xi: xtmixed glucose i.day || dayXrun: x, variance nolog
```

4.6 Nested and crossed factors experimental design

Here I demonstrate how `xtmixed` can be used to fit random-effects design with crossed and nested factors. We simulate data from the following experiment. Ten measurements ($r = 10$) are obtained for each of the three machines ($a = 3$) from a random sample of three runs ($c = 3$) for 3 days ($b = 3$). Runs are nested within day, and machines are crossed with runs and days. Machine effect is a fixed effect, and all other effects are random. The variance components for this design are variability among days ($\sigma_b^2 = 2.25$), variability among runs within day ($\sigma_{c(b)}^2 = 0.09$), variability due to the interaction between machine and day ($\sigma_{ab}^2 = 0.25$), and variability due to the interaction between machine and runs nested within day ($\sigma_{ac(b)}^2 = 0.64$); the error variance is set to one ($\sigma_e^2 = 1$).

We use the following syntax for `anova` to produce a table corresponding to this design:

```
. use simul
(Simulation: crossed and nested factors)
. anova measurement machine / day machine*day run|day/ machine*run|day
                           Number of obs =     270     R-squared     =  0.8952
                           Root MSE      =  .97025     Adj R-squared =  0.8840
```

Source	Partial SS	df	MS	F	Prob > F
Model	1953.37309	26	75.1297341	79.81	0.0000
machine	1192.8791	2	596.439552	2.29	0.3044
day	521.966092	2	260.983046		
machine*day	72.3036746	4	18.0759186	2.22	0.1281
run|day	68.4783724	6	11.4130621	1.40	0.2909
machine*run|day	97.7458432	12	8.14548694		
Residual	228.756698	243	.941385588		
Total	2182.12978	269	8.11200663		

Using the following formulas, we can estimate variance components from the above `anova` table

$$\widehat{\sigma}^2_e = \mathrm{MS}(\texttt{Residual})$$

$$\widehat{\sigma}^2_{ac(b)} = \frac{\mathrm{MS}(\texttt{machine*run|day}) - \mathrm{MS}(\texttt{Residual})}{r}$$

$$\widehat{\sigma}^2_{c(b)} = \frac{\mathrm{MS}(\texttt{machine*run|day}) - \mathrm{MS}(\texttt{run|day})}{ar}$$

$$\widehat{\sigma}^2_{ab} = \frac{\mathrm{MS}(\texttt{machine*run|day}) - \mathrm{MS}(\texttt{machine*day})}{cr}$$

$$\widehat{\sigma}^2_{b} = \frac{\mathrm{MS}(\texttt{day}) + \mathrm{MS}(\texttt{machine*run|day}) - \mathrm{MS}(\texttt{machine*day}) - \mathrm{MS}(\texttt{run|day})}{acr}$$

to be $\widehat{\sigma}^2_e = 0.94$, $\widehat{\sigma}^2_{ac(b)} = 0.72$, $\widehat{\sigma}^2_{c(b)} = 0.109$, $\widehat{\sigma}^2_{ab} = 0.33$, and $\widehat{\sigma}^2_b = 2.66$.

Let us first show the brute-force way of fitting `xtmixed` for this random-effects design. We first need to create all corresponding interaction terms using `egen`:

```
. egen dayXrun = group(day run)
. egen machXday = group(machine day)
. egen machXdayXrun = group(machine day run)
```

```
. xi: xtmixed measurement i.machine || _all: R.day || _all: R.dayXrun || _all:
> R.machXday || _all: R.machXdayXrun, variance nolog
i.machine         _Imachine_1-3       (naturally coded; _Imachine_1 omitted)

Mixed-effects REML regression                   Number of obs      =       270
Group variable: _all                            Number of groups   =         1

                                                Obs per group: min =       270
                                                               avg =     270.0
                                                               max =       270

                                                Wald chi2(2)       =     65.99
Log restricted-likelihood = -409.51008          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
 measurement |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Imachine_2 |   1.539087   .6337863     2.43   0.015     .2968884    2.781285
  _Imachine_3 |   5.024508   .6337863     7.93   0.000      3.78231    6.266707
        _cons |  -.0387583   1.049048    -0.04   0.971    -2.094854    2.017337
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
_all: Identity               |
                  var(R.day) |    2.66267   2.904453      .3139226    22.58458
-----------------------------+------------------------------------------------
_all: Identity               |
              var(R.dayXrun) |   .1089146   .2460279      .0013011    9.117384
-----------------------------+------------------------------------------------
_all: Identity               |
             var(R.machXday) |   .3310101   .4402353      .0244211    4.486609
-----------------------------+------------------------------------------------
_all: Identity               |
             var(R.machXd~n) |   .7204136   .3326501      .2914342    1.780833
-----------------------------+------------------------------------------------
               var(Residual) |   .9413857   .0854042      .7880341    1.124579
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(4) =   301.88   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference
```

Now I demonstrate the more efficient way of using `xtmixed` to fit this design. The interaction terms `machXday` and `machXdayXrun` can be viewed as the following nested terms: machine nested within days and runs nested within machines nested within days. Therefore, we have three levels of nesting, with `day` defining the first level, `machine` defining the second level, and `run` defining the third level. This formulation allows us to obtain variance components σ_b^2, σ_{ab}^2, and $\sigma_{ac(b)}^2$, respectively. To obtain the variance component, $\sigma_{c(b)}^2$, recall that we can obtain the estimate of a variance component for a nested factor by using the exchangeable covariance matrix as described in section 4.4. All the above suggest the following syntax for `xtmixed`:

(*Continued on next page*)

```
. xi: xtmixed measurement i.machine || day: R.run, cov(exchangeable) || machine
> : || run:, variance nolog
i.machine         _Imachine_1-3       (naturally coded; _Imachine_1 omitted)
Mixed-effects REML regression                   Number of obs      =       270

-----------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+------------------------------------------
            day |        3         90       90.0         90
        machine |        9         30       30.0         30
            run |       27         10       10.0         10
-----------------------------------------------------------

                                                Wald chi2(2)       =     65.99
Log restricted-likelihood = -409.51008          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
 measurement |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Imachine_2 |   1.539087   .6337875     2.43   0.015      .296886    2.781287
 _Imachine_3 |   5.024508   .6337875     7.93   0.000     3.782308    6.266709
       _cons |  -.0387583   1.049048    -0.04   0.971    -2.094855    2.017339
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
day: Exchangeable            |
                  var(R.run) |   2.771589   2.907932       .354531    21.66724
                  cov(R.run) |   2.662671    2.90446     -3.029966    8.355309
-----------------------------+------------------------------------------------
machine: Identity            |
                  var(_cons) |   .3310135   .4402365      .0244218    4.486568
-----------------------------+------------------------------------------------
run: Identity                |
                  var(_cons) |   .7204107   .3326466      .2914349    1.780815
-----------------------------+------------------------------------------------
               var(Residual) |   .9413856   .0854042       .788034    1.124579
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(4) =   301.88   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference
```

5 Summary

In this article, I described how variance components can be obtained in Stata with the emphasis on using `xtmixed`. I demonstrated effective ways of fitting different ANOVA models with `xtmixed` by expressing them as multilevel models, also providing the syntax corresponding to the direct translation of the ANOVA model. The latter model provides a straightforward approach for fitting random-effects designs with `xtmixed` by directly constructing the design matrix for random effects. With the former, however, there are no general rules for reexpressing a generic random-effects design as a multilevel model. Trial and error may be required to find the most efficient way to fit random-effects designs with `xtmixed`.

Stata users are advised to use the alternate multilevel formulation for random-effects designs with many levels of random effects. It may be difficult for certain designs to find the same formulation as a multilevel model, and the direct way of fitting may be infeasible because of the large number of levels. In such situations, you might obtain results by using the brute-force approach on a subset of data with fewer levels and then find the multilevel representation that matches your results. Then this formulation can be used to fit the model on the entire dataset.

Although the examples considered in this article correspond to balanced designs, `xtmixed` can also be used with unbalanced designs.

6 References

Goldstein, H. 2003. *Multilevel Statistical Models*. 3rd ed. London: Arnold.

Hartley, H. O., and J. N. K. Rao. 1967. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* 54: 93–108.

Henderson, C. R. 1953. Estimation of variance and covariance components. *Biometrics* 9: 226–252.

Kuehl, R. O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*. 2nd ed. Pacific Grove, CA: Brooks/Cole.

Rabe-Hesketh, S., and A. Skrondal. 2005. *Multilevel and Longitudinal Modeling Using Stata*. College Station, TX: Stata Press.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. *Variance Components*. New York: Wiley.

Thompson, S. K. 1992. *Sampling*. New York: Wiley.

Tippett, L. H. C. 1931. *The Methods of Statistics*. London: Williams and Norgate.

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

About the author

Yulia Marchenko is a statistician at StataCorp.