### Syntax

```stata
mixed depvar fe_equation [ || re_equation] [ || re_equation ...] [, options]
```

where the syntax of `fe_equation` is

```stata
[ indepvars ] [ if ] [ in ] [ weight ] [, fe_options ]
```

and the syntax of `re_equation` is one of the following:

- for random coefficients and intercepts
  ```stata
  levelvar: [ varlist ] [, re_options ]
  ```

- for random effects among the values of a factor variable
  ```stata
  levelvar: R. varname [, re_options ]
  ```

`levelvar` is a variable identifying the group structure for the random effects at that level or is `_all` representing one group comprising all observations.

### Description

<table>
<thead>
<tr>
<th>fe_options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>noconstant</code></td>
<td>suppress constant term from the fixed-effects equation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>re_options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>covariance(vartype)</code></td>
<td>variance–covariance structure of the random effects</td>
</tr>
<tr>
<td><code>noconstant</code></td>
<td>suppress constant term from the random-effects equation</td>
</tr>
<tr>
<td><code>collinear</code></td>
<td>keep collinear variables</td>
</tr>
<tr>
<td><code>fweight(exp)</code></td>
<td>frequency weights at higher levels</td>
</tr>
<tr>
<td><code>pweight(exp)</code></td>
<td>sampling weights at higher levels</td>
</tr>
</tbody>
</table>
## mixed — Multilevel mixed-effects linear regression

<table>
<thead>
<tr>
<th>options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model</strong></td>
<td></td>
</tr>
<tr>
<td><code>ml</code></td>
<td>fit model via maximum likelihood; the default</td>
</tr>
<tr>
<td><code>reml</code></td>
<td>fit model via restricted maximum likelihood</td>
</tr>
<tr>
<td><code>pwscale(scale_method)</code></td>
<td>control scaling of sampling weights in two-level models</td>
</tr>
<tr>
<td><code>residuals(rspec)</code></td>
<td>structure of residual errors</td>
</tr>
<tr>
<td><strong>SE/Robust</strong></td>
<td></td>
</tr>
<tr>
<td><code>vce(vcetype)</code></td>
<td><code>vcetype</code> may be <code>oim</code>, <code>robust</code>, or <code>cluster clustvar</code></td>
</tr>
<tr>
<td><strong>Reporting</strong></td>
<td></td>
</tr>
<tr>
<td><code>level(#)</code></td>
<td>set confidence level; default is <code>level(95)</code></td>
</tr>
<tr>
<td><code>variance</code></td>
<td>show random-effects and residual-error parameter estimates as variances and covariances; the default</td>
</tr>
<tr>
<td><code>stddeviations</code></td>
<td>show random-effects and residual-error parameter estimates as standard deviations</td>
</tr>
<tr>
<td><code>noretable</code></td>
<td>suppress random-effects table</td>
</tr>
<tr>
<td><code>nofetable</code></td>
<td>suppress fixed-effects table</td>
</tr>
<tr>
<td><code>estmetric</code></td>
<td>show parameter estimates in the estimation metric</td>
</tr>
<tr>
<td><code>noheader</code></td>
<td>suppress output header</td>
</tr>
<tr>
<td><code>nogroup</code></td>
<td>suppress table summarizing groups</td>
</tr>
<tr>
<td><code>nostderr</code></td>
<td>do not estimate standard errors of random-effects parameters</td>
</tr>
<tr>
<td><code>nolrtest</code></td>
<td>do not perform likelihood-ratio test comparing with linear regression</td>
</tr>
<tr>
<td><code>display_options</code></td>
<td>control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling</td>
</tr>
<tr>
<td><strong>EM options</strong></td>
<td></td>
</tr>
<tr>
<td><code>emiterate(#)</code></td>
<td>number of EM iterations; default is <code>emiterate(20)</code></td>
</tr>
<tr>
<td><code>emtolerance(#)</code></td>
<td>EM convergence tolerance; default is <code>emtolerance(1e-10)</code></td>
</tr>
<tr>
<td><code>emonly</code></td>
<td>fit model exclusively using EM</td>
</tr>
<tr>
<td><code>emlog</code></td>
<td>show EM iteration log</td>
</tr>
<tr>
<td><code>emdots</code></td>
<td>show EM iterations as dots</td>
</tr>
<tr>
<td><strong>Maximization</strong></td>
<td></td>
</tr>
<tr>
<td><code>maximize_options</code></td>
<td>control the maximization process; seldom used</td>
</tr>
<tr>
<td><code>matsqrt</code></td>
<td>parameterize variance components using matrix square roots; the default</td>
</tr>
<tr>
<td><code>matlog</code></td>
<td>parameterize variance components using matrix logarithms</td>
</tr>
<tr>
<td><code>coeflegend</code></td>
<td>display legend instead of statistics</td>
</tr>
</tbody>
</table>
mixed — Multilevel mixed-effects linear regression

vartype

<table>
<thead>
<tr>
<th>vartype</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>independent</td>
<td>one unique variance parameter per random effect, all covariances 0; the default unless the R. notation is used</td>
</tr>
<tr>
<td>exchangeable</td>
<td>equal variances for random effects, and one common pairwise covariance</td>
</tr>
<tr>
<td>identity</td>
<td>equal variances for random effects, all covariances 0; the default if the R. notation is used</td>
</tr>
<tr>
<td>unstructured</td>
<td>all variances and covariances to be distinctly estimated</td>
</tr>
</tbody>
</table>

indepvars may contain factor variables; see [U] 11.4.3 Factor variables.

depvar, indepvars, and varlist may contain time-series operators; see [U] 11.4.4 Time-series varlists.

bootstrap, by, jackknife, mi estimate, rolling, and statsby are allowed; see [U] 11.1.10 Prefix commands.

Weights are not allowed with the bootstrap prefix; see [R] bootstrap.
pweights and fweights are allowed; see [U] 11.1.6 weight.

coefflegend does not appear in the dialog box.

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

Menu

Statistics > Multilevel mixed-effects models > Linear regression

Description

mixed fits linear mixed-effects models. The overall error distribution of the linear mixed-effects model is assumed to be Gaussian, and heteroskedasticity and correlations within lowest-level groups also may be modeled.

Options

noconstant suppresses the constant (intercept) term and may be specified for the fixed-effects equation and for any or all of the random-effects equations.

covariance(vartype), where vartype is

\[
\text{independent} \mid \text{exchangeable} \mid \text{identity} \mid \text{unstructured}
\]

specifies the structure of the covariance matrix for the random effects and may be specified for each random-effects equation. An independent covariance structure allows for a distinct variance for each random effect within a random-effects equation and assumes that all covariances are 0. exchangeable structure specifies one common variance for all random effects and one common pairwise covariance. identity is short for “multiple of the identity”; that is, all variances are equal and all covariances are 0. unstructured allows for all variances and covariances to be distinct. If an equation consists of \( p \) random-effects terms, the unstructured covariance matrix will have \( p(p + 1)/2 \) unique parameters.

covariance(independent) is the default, except when the R. notation is used, in which case covariance(identity) is the default and only covariance(identity) and covariance(exchangeable) are allowed.
**mixed** specifies that **mixed** not omit collinear variables from the random-effects equation. Usually, there is no reason to leave collinear variables in place; in fact, doing so usually causes the estimation to fail because of the matrix singularity caused by the collinearity. However, with certain models (for example, a random-effects model with a full set of contrasts), the variables may be collinear, yet the model is fully identified because of restrictions on the random-effects covariance structure. In such cases, using the **collinear** option allows the estimation to take place with the random-effects equation intact.

**fweight**(*exp*) specifies frequency weights at higher levels in a multilevel model, whereas frequency weights at the first level (the observation level) are specified in the usual manner, for example, [fw=fwtvar1]. *exp* can be any valid Stata expression, and you can specify **fweight()** at levels two and higher of a multilevel model. For example, in the two-level model

```stata
.mixed fixed_portion [fw = wt1] || school: ..., fweight(wt2) ...
```

the variable wt1 would hold the first-level (the observation-level) frequency weights, and wt2 would hold the second-level (the school-level) frequency weights.

**pweight**(*exp*) specifies sampling weights at higher levels in a multilevel model, whereas sampling weights at the first level (the observation level) are specified in the usual manner, for example, [pw=pwtvar1]. *exp* can be any valid Stata expression, and you can specify **pweight()** at levels two and higher of a multilevel model. For example, in the two-level model

```stata
.mixed fixed_portion [pw = wt1] || school: ..., pweight(wt2) ...
```

variable wt1 would hold the first-level (the observation-level) sampling weights, and wt2 would hold the second-level (the school-level) sampling weights.

See *Survey data* in Remarks and examples below for more information regarding the use of sampling weights in multilevel models.

Weighted estimation, whether frequency or sampling, is not supported under restricted maximum-likelihood estimation (REML).

**mle** and **reml** specify the statistical method for fitting the model.

- **mle**, the default, specifies that the model be fit using maximum likelihood (ML).
- **reml** specifies that the model be fit using restricted maximum likelihood (REML), also known as residual maximum likelihood.

**pwscale**(scale_method), where scale_method is

```stata
| size | effective | gk |
```

controls how sampling weights (if specified) are scaled in two-level models.

- **scale_method size** specifies that first-level (observation-level) weights be scaled so that they sum to the sample size of their corresponding second-level cluster. Second-level sampling weights are left unchanged.

- **scale_method effective** specifies that first-level weights be scaled so that they sum to the effective sample size of their corresponding second-level cluster. Second-level sampling weights are left unchanged.

- **scale_method gk** specifies the Graubard and Korn (1996) method. Under this method, second-level weights are set to the cluster averages of the products of the weights at both levels, and first-level weights are then set equal to 1.

**pwscale()** is supported only with two-level models. See *Survey data* in Remarks and examples below for more details on using **pwscale()**.
residuals(spec), where spec is

\[
\text{restype} \ [ \ , \ \text{residual\_options} ]
\]

specifies the structure of the residual errors within the lowest-level groups (the second level of a multilevel model with the observations comprising the first level) of the linear mixed model. For example, if you are modeling random effects for classes nested within schools, then residuals() refers to the residual variance–covariance structure of the observations within classes, the lowest-level groups.

restype is

\[
\text{independent} | \text{exchangeable} | \text{ar} \ # | \text{ma} \ # | \text{unstructured}\]

\[
\text{banded} \ # | \text{toeplitz} \ # | \text{exponential}
\]

By default, restype is independent, which means that all residuals are independent and identically distributed (i.i.d.) Gaussian with one common variance. When combined with by(varname), independence is still assumed, but you estimate a distinct variance for each level of varname. Unlike with the structures described below, varname does not need to be constant within groups.

restype exchangeable estimates two parameters, one common within-group variance and one common pairwise covariance. When combined with by(varname), these two parameters are distinctly estimated for each level of varname. Because you are modeling a within-group covariance, varname must be constant within lowest-level groups.

restype ar # assumes that within-group errors have an autoregressive (AR) structure of order #; ar 1 is the default. The t(varname) option is required, where varname is an integer-valued time variable used to order the observations within groups and to determine the lags between successive observations. Any nonconsecutive time values will be treated as gaps. For this structure, # + 1 parameters are estimated (# AR coefficients and one overall error variance). restype ar may be combined with by(varname), but varname must be constant within groups.

restype ma # assumes that within-group errors have a moving average (MA) structure of order #; ma 1 is the default. The t(varname) option is required, where varname is an integer-valued time variable used to order the observations within groups and to determine the lags between successive observations. Any nonconsecutive time values will be treated as gaps. For this structure, # + 1 parameters are estimated (# MA coefficients and one overall error variance). restype ma may be combined with by(varname), but varname must be constant within groups.

restype unstructured is the most general structure; it estimates distinct variances for each within-group error and distinct covariances for each within-group error pair. The t(varname) option is required, where varname is a nonnegative-integer–valued variable that identifies the observations within each group. The groups may be unbalanced in that not all levels of t() need to be observed within every group, but you may not have repeated t() values within any particular group. When you have p levels of t(), then \( p(p+1)/2 \) parameters are estimated. restype unstructured may be combined with by(varname), but varname must be constant within groups.

restype banded # is a special case of unstructured that restricts estimation to the covariances within the first # off-diagonals and sets the covariances outside this band to 0. The t(varname) option is required, where varname is a nonnegative-integer–valued variable that identifies the observations within each group. # is an integer between 0 and \( p-1 \), where \( p \) is the number of levels of t(). By default, # is \( p-1 \); that is, all elements
of the covariance matrix are estimated. When \( \# \) is 0, only the diagonal elements of the covariance matrix are estimated. `restype banded` may be combined with `by(varname)`, but `varname` must be constant within groups.

`restype toeplitz \#` assumes that within-group errors have Toeplitz structure of order \( \# \), for which correlations are constant with respect to time lags less than or equal to \( \# \) and are 0 for lags greater than \( \# \). The `t(varname)` option is required, where `varname` is an integer-valued time variable used to order the observations within groups and to determine the lags between successive observations. \( \# \) is an integer between 1 and the maximum observed lag (the default). Any nonconsecutive time values will be treated as gaps. For this structure, \( \# + 1 \) parameters are estimated (\( \# \) correlations and one overall error variance). `restype toeplitz` may be combined with `by(varname)`, but `varname` must be constant within groups.

`restype exponential` is a generalization of the AR covariance model that allows for unequally spaced and noninteger time values. The `t(varname)` option is required, where `varname` is real-valued. For the exponential covariance model, the correlation between two errors is the parameter \( \rho \), raised to a power equal to the absolute value of the difference between the `t()` values for those errors. For this structure, two parameters are estimated (the correlation parameter \( \rho \) and one overall error variance). `restype exponential` may be combined with `by(varname)`, but `varname` must be constant within groups.

`residual_options` are `by(varname)` and `t(varname)`.

`by(varname)` is for use within the `residuals()` option and specifies that a set of distinct residual-error parameters be estimated for each level of `varname`. In other words, you use `by()` to model heteroskedasticity.

`t(varname)` is for use within the `residuals()` option to specify a time variable for the `ar`, `ma`, `toeplitz`, and `exponential` structures, or to identify the observations when `restype` is unstructured or banded.

`vce(vcetype)` specifies the type of standard error reported, which includes types that are derived from asymptotic theory (`oim`), that are robust to some kinds of misspecification (`robust`), and that allow for intragroup correlation (`cluster clustvar`); see [R] `vce_option`. If `vce(robust)` is specified, robust variances are clustered at the highest level in the multilevel model. `vce(robust)` and `vce(cluster clustvar)` are not supported with REML estimation.

`level(#);` see [R] `estimation options`.

`variance`, the default, displays the random-effects and residual-error parameter estimates as variances and covariances.

`stddeviations` displays the random-effects and residual-error parameter estimates as standard deviations and correlations.

`noretetable` suppresses the random-effects table from the output.

`nofetable` suppresses the fixed-effects table from the output.

`estmetric` displays all parameter estimates in the estimation metric. Fixed-effects estimates are unchanged from those normally displayed, but random-effects parameter estimates are displayed as log-standard deviations and hyperbolic arctangents of correlations, with equation names that
organize them by model level. Residual-variance parameter estimates are also displayed in their original estimation metric.

noheader suppresses the output header, either at estimation or upon replay.

nogroup suppresses the display of group summary information (number of groups, average group size, minimum, and maximum) from the output header.

nostderr prevents `mixed` from calculating standard errors for the estimated random-effects parameters, although standard errors are still provided for the fixed-effects parameters. Specifying this option will speed up computation times. `nostderr` is available only when residuals are modeled as independent with constant variance.

nolrtest prevents `mixed` from fitting a reference linear regression model and using this model to calculate a likelihood-ratio test comparing the mixed model to ordinary regression. This option may also be specified on replay to suppress this test from the output.

`display_options`: `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(style)`, `cformat(\%fmt)`, `pformat(\%fmt)`, `sformat(\%fmt)`, and `nolstretch`; see [R] estimation options.

---

**EM options**

These options control the expectation-maximization (EM) iterations that take place before estimation switches to a gradient-based method. When residuals are modeled as independent with constant variance, EM will either converge to the solution or bring parameter estimates close to the solution. For other residual structures or for weighted estimation, EM is used to obtain starting values.

`emiterate(#)` specifies the number of EM iterations to perform. The default is `emiterate(20)`.

`emtolerance(#)` specifies the convergence tolerance for the EM algorithm. The default is `emtolerance(1e-10)`. EM iterations will be halted once the log (restricted) likelihood changes by a relative amount less than #. At that point, optimization switches to a gradient-based method, unless `emonly` is specified, in which case maximization stops.

`emonly` specifies that the likelihood be maximized exclusively using EM. The advantage of specifying `emonly` is that EM iterations are typically much faster than those for gradient-based methods. The disadvantages are that EM iterations can be slow to converge (if at all) and that EM provides no facility for estimating standard errors for the random-effects parameters. `emonly` is available only with unweighted estimation and when residuals are modeled as independent with constant variance.

`emlog` specifies that the EM iteration log be shown. The EM iteration log is, by default, not displayed unless the `emonly` option is specified.

`emdots` specifies that the EM iterations be shown as dots. This option can be convenient because the EM algorithm may require many iterations to converge.

---

**Maximization**

`maximize_options`: `difficult`, `technique(algorithm_spec)`, `iterate(#)`, `[no]log`, `trace`, `gradient`, `showstep`, `hessian`, `showtolerance`, `tolerance(#)`, `ltolerance(#)`, `nrtolerance(#)`, and `nonrtolerance`; see [R] maximize. Those that require special mention for `mixed` are listed below.

For the `technique()` option, the default is `technique(nr)`. The bhhh algorithm may not be specified.

`matsqrt` (the default), during optimization, parameterizes variance components by using the matrix square roots of the variance–covariance matrices formed by these components at each model level.
matlog, during optimization, parameterizes variance components by using the matrix logarithms of the variance–covariance matrices formed by these components at each model level.

The matsqrt parameterization ensures that variance–covariance matrices are positive semidefinite, while matlog ensures matrices that are positive definite. For most problems, the matrix square root is more stable near the boundary of the parameter space. However, if convergence is problematic, one option may be to try the alternate matlog parameterization. When convergence is not an issue, both parameterizations yield equivalent results.

The following option is available with mixed but is not shown in the dialog box: coeflegend; see [R] estimation options.

### Remarks and examples

Remarks are presented under the following headings:

**Introduction**

**Two-level models**

**Covariance structures**

**Likelihood versus restricted likelihood**

**Three-level models**

**Blocked-diagonal covariance structures**

**Heteroskedastic random effects**

**Heteroskedastic residual errors**

**Other residual-error structures**

**Crossed-effects models**

**Diagnosing convergence problems**

**Survey data**

### Introduction

Linear mixed models are models containing both fixed effects and random effects. They are a generalization of linear regression allowing for the inclusion of random deviations (effects) other than those associated with the overall error term. In matrix notation,

\[
\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{\epsilon}
\]

where \(\mathbf{y}\) is the \(n \times 1\) vector of responses, \(\mathbf{X}\) is an \(n \times p\) design/covariate matrix for the fixed effects \(\beta\), and \(\mathbf{Z}\) is the \(n \times q\) design/covariate matrix for the random effects \(\mathbf{u}\). The \(n \times 1\) vector of errors \(\mathbf{\epsilon}\) is assumed to be multivariate normal with mean 0 and variance matrix \(\sigma^2_{\epsilon}\mathbf{R}\).

The fixed portion of (1), \(\mathbf{X}\beta\), is analogous to the linear predictor from a standard OLS regression model with \(\beta\) being the regression coefficients to be estimated. For the random portion of (1), \(\mathbf{Z}\mathbf{u} + \mathbf{\epsilon}\), we assume that \(\mathbf{u}\) has variance–covariance matrix \(\mathbf{G}\) and that \(\mathbf{u}\) is orthogonal to \(\mathbf{\epsilon}\) so that

\[
\text{Var} \left[ \begin{bmatrix} \mathbf{u} \\ \mathbf{\epsilon} \end{bmatrix} \right] = \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \sigma^2_{\epsilon}\mathbf{R} \end{bmatrix}
\]

The random effects \(\mathbf{u}\) are not directly estimated (although they may be predicted), but instead are characterized by the elements of \(\mathbf{G}\), known as variance components, that are estimated along with the overall residual variance \(\sigma^2_{\epsilon}\) and the residual-variance parameters that are contained within \(\mathbf{R}\).
The general forms of the design matrices $X$ and $Z$ allow estimation for a broad class of linear models: blocked designs, split-plot designs, growth curves, multilevel or hierarchical designs, etc. They also allow a flexible method of modeling within-cluster correlation. Subjects within the same cluster can be correlated as a result of a shared random intercept, or through a shared random slope on (say) age, or both. The general specification of $G$ also provides additional flexibility—the random intercept and random slope could themselves be modeled as independent, or correlated, or independent with equal variances, and so forth. The general structure of $R$ also allows for residual errors to be heteroskedastic and correlated, and allows flexibility in exactly how these characteristics can be modeled.

Comprehensive treatments of mixed models are provided by, among others, Searle, Casella, and McCulloch (1992); McCulloch, Searle, and Neuhaus (2008); Verbeke and Molenberghs (2000); Raudenbush and Bryk (2002); Demidenko (2004); and Pinheiro and Bates (2000). In particular, chapter 2 of Searle, Casella, and McCulloch (1992) provides an excellent history.

The key to fitting mixed models lies in estimating the variance components, and for that there exist many methods. Most of the early literature in mixed models dealt with estimating variance components in ANOVA models. For simple models with balanced data, estimating variance components amounts to solving a system of equations obtained by setting expected mean-squares expressions equal to their observed counterparts. Much of the work in extending the ANOVA method to unbalanced data for general ANOVA designs is due to Henderson (1953).

The ANOVA method, however, has its shortcomings. Among these is a lack of uniqueness in that alternative, unbiased estimates of variance components could be derived using other quadratic forms of the data in place of observed mean squares (Searle, Casella, and McCulloch 1992, 38–39). As a result, ANOVA methods gave way to more modern methods, such as minimum norm quadratic unbiased estimation (MINQUE) and minimum variance quadratic unbiased estimation (MIVQUE); see Rao (1973) for MINQUE and LaMotte (1973) for MIVQUE. Both methods involve finding optimal quadratic forms of the data that are unbiased for the variance components.

The most popular methods, however, are ML and REML, and these are the two methods that are supported by 

\[

y_j = X_j \beta + Z_j u_j + \epsilon_j

\]

for $j = 1, \ldots, M$, with cluster $j$ consisting of $n_j$ observations. The response $y_j$ comprises the rows of $y$ corresponding with the $j$th cluster, with $X_j$ and $\epsilon_j$ defined analogously. The random effects $u_j$ can now be thought of as $M$ realizations of a $q \times 1$ vector that is normally distributed with mean 0 and $q \times q$ variance matrix $\Sigma$. The matrix $Z_i$ is the $n_j \times q$ design matrix for the $j$th cluster random effects. Relating this to (1), note that

\[
Z = \begin{bmatrix}
Z_1 & 0 & \cdots & 0 \\
0 & Z_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & Z_M
\end{bmatrix} ; 
\quad u = \begin{bmatrix}
u_1 \\
\vdots \\
u_M
\end{bmatrix} ; 
\quad G = I_M \otimes \Sigma ; 
\quad R = I_M \otimes \Lambda
\]

The mixed-model formulation (2) is from Laird and Ware (1982) and offers two key advantages. First, it makes specifications of random-effects terms easier. If the clusters are schools, you can
simply specify a random effect at the school level, as opposed to thinking of what a school-level random effect would mean when all the data are considered as a whole (if it helps, think Kronecker products). Second, representing a mixed-model with (2) generalizes easily to more than one set of random effects. For example, if classes are nested within schools, then (2) can be generalized to allow random effects at both the school and the class-within-school levels. This we demonstrate later.

In the sections that follow, we assume that residuals are independent with constant variance; that is, in (3) we treat $\Lambda$ equal to the identity matrix and limit ourselves to estimating one overall residual variance, $\sigma^2_e$. Beginning in *Heteroskedastic residual errors*, we relax this assumption.

**Two-level models**

We begin with a simple application of (2) as a two-level model, because a one-level linear model, by our terminology, is just standard OLS regression.

> **Example 1**

Consider a longitudinal dataset, used by both Ruppert, Wand, and Carroll (2003) and Diggle et al. (2002), consisting of weight measurements of 48 pigs on 9 successive weeks. Pigs are identified by the variable `id`. Below is a plot of the growth curves for the first 10 pigs.

```
. use http://www.stata-press.com/data/r13/pig
   (Longitudinal analysis of pig weights)
. twoway connected weight week if id<=10, connect(L)
```

It seems clear that each pig experiences a linear trend in growth and that overall weight measurements vary from pig to pig. Because we are not really interested in these particular 48 pigs per se, we instead treat them as a random sample from a larger population and model the between-pig variability as a random effect, or in the terminology of (2), as a random-intercept term at the pig level. We thus wish to fit the model

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{week}_{ij} + u_j + \epsilon_{ij}$$

for $i = 1, \ldots, 9$ weeks and $j = 1, \ldots, 48$ pigs. The fixed portion of the model, $\beta_0 + \beta_1 \text{week}_{ij}$, simply states that we want one overall regression line representing the population average. The random effect $u_j$ serves to shift this regression line up or down according to each pig. Because the random effects occur at the pig level (`id`), we fit the model by typing
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0:  log likelihood = -1014.9268
Iteration 1:  log likelihood = -1014.9268
Computing standard errors:
Mixed-effects ML regression  Number of obs = 432
Group variable: id  Number of groups = 48
Obs per group: min = 9
              avg = 9.0
              max = 9
Wald chi2(1) = 25337.49
Log likelihood = -1014.9268  Prob > chi2 = 0.0000

| weight | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval]         |
|--------|--------|-----------|-------|------|-----------------------------|
| week   | 6.209896 | 0.0390124 | 159.18 | 0.000 | 6.133433 6.286359           |
| _cons  | 19.35561 | 0.5974059 | 32.40  | 0.000 | 18.18472 20.52651           |

Random-effects Parameters  Estimate  Std. Err.  [95% Conf. Interval]
| id: Identity | var(_cons) | 14.81751 | 3.124226 | 9.801716 22.40002 |
| var(Residual) | 4.383264 | 0.3163348 | 3.805112 5.04926 |

LR test vs. linear regression: chibar2(01) = 472.65 Prob >= chibar2 = 0.0000

Notes:

1. By typing `weight week || id;`, we specified the response, `weight`, and the fixed portion of the model in the same way that we would if we were using `regress` or any other estimation command. Our fixed effects are a coefficient on `week` and a constant term.

2. When we added `|| id;`, we specified random effects at the level identified by the group variable `id`, that is, the pig level (level two). Because we wanted only a random intercept, that is all we had to type.

3. The estimation log consists of three parts:
   a. A set of EM iterations used to refine starting values. By default, the iterations themselves are not displayed, but you can display them with the `emlog` option.
   b. A set of gradient-based iterations. By default, these are Newton–Raphson iterations, but other methods are available by specifying the appropriate `maximize` options; see `[R] maximize`.
   c. The message “Computing standard errors”. This is just to inform you that `mixed` has finished its iterative maximization and is now reparameterizing from a matrix-based parameterization (see `Methods and formulas`) to the natural metric of variance components and their estimated standard errors.

4. The output title, “Mixed-effects ML regression”, informs us that our model was fit using ML, the default. For REML estimates, use the `reml` option.

Because this model is a simple random-intercept model fit by ML, it would be equivalent to using `xtreg` with its `mle` option.

5. The first estimation table reports the fixed effects. We estimate $\beta_0 = 19.36$ and $\beta_1 = 6.21$. 
6. The second estimation table shows the estimated variance components. The first section of the table is labeled \textit{id: Identity}, meaning that these are random effects at the \textit{id} (pig) level and that their variance–covariance matrix is a multiple of the identity matrix; that is, \( \Sigma = \sigma^2_u \mathbf{I} \). Because we have only one random effect at this level, \texttt{mixed} knew that \textit{Identity} is the only possible covariance structure. In any case, the variance of the level-two errors, \( \sigma^2_u \), is estimated as 14.82 with standard error 3.12.

7. The row labeled \textit{var(Residual)} displays the estimated variance of the overall error term; that is, \( \hat{\sigma}_\epsilon^2 = 4.38 \). This is the variance of the level-one errors, that is, the residuals.

8. Finally, a likelihood-ratio test comparing the model with one-level ordinary linear regression, model (4) without \( u_j \), is provided and is highly significant for these data.

We now store our estimates for later use:

\[
\text{. estimates store randint}
\]

\textbf{Example 2}

Extending (4) to allow for a random slope on \textit{week} yields the model

\[
\text{weight}_{ij} = \beta_0 + \beta_1 \text{week}_{ij} + u_{0j} + u_{1j} \text{week}_{ij} + \epsilon_{ij}
\]

(5)

and we fit this with \texttt{mixed}:

\[
\text{mixed weight week || id: week}
\]

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -869.03825
Iteration 1: log likelihood = -869.03825

Computing standard errors:

Mixed-effects ML regression

| Number of obs | 432 |
| Number of groups | 48 |

Obs per group: min = 9
avg = 9.0
max = 9

Wald chi2(1) = 4689.51

Log likelihood = -869.03825

Prob > chi2 = 0.0000

\[
\begin{array}{cccccc}
\text{weight} & \text{Coeff.} & \text{Std. Err.} & z & P>|z| & [95\% \text{ Conf. Interval}] \\
\text{week} & 6.209896 & .0906819 & 68.48 & 0.000 & 6.032163 \text{ -- 6.387629} \\
_{\text{cons}} & 19.35561 & .3979159 & 48.64 & 0.000 & 18.57571 \text{ -- 20.13551} \\
\end{array}
\]

\[
\begin{array}{ccccc}
\text{Random-effects Parameters} & \text{Estimate} & \text{Std. Err.} & [95\% \text{ Conf. Interval}] \\
\text{id: Independent} & & & & \\
\text{var(week)} & .3680668 & .0801181 & .2402389 \text{ -- 5.639103} \\
\text{var(_cons)} & 6.756364 & 1.543503 & 4.317721 \text{ -- 10.57235} \\
\text{var(Residual)} & 1.598811 & .1233988 & 1.374358 \text{ -- 1.85992} \\
\end{array}
\]

LR test vs. linear regression: \[ \text{\texttt{ch}i2(2) = 764.42 \ Prob > \text{\texttt{ch}i2} = 0.0000} \]

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

\[
\text{. estimates store randslope}
\]
Because we did not specify a covariance structure for the random effects \((u_{0j}, u_{1j})'\), mixed used the default Independent structure; that is,

\[
\Sigma = \text{Var} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \sigma_{u0}^2 & 0 \\ 0 & \sigma_{u1}^2 \end{bmatrix}
\]

(6)

with \(\hat{\sigma}_{u0}^2 = 6.76\) and \(\hat{\sigma}_{u1}^2 = 0.37\). Our point estimates of the fixed effects are essentially identical to those from model (4), but note that this does not hold generally. Given the 95% confidence interval for \(\hat{\sigma}_{u1}^2\), it would seem that the random slope is significant, and we can use lrtest and our two stored estimation results to verify this fact:

\[
. \text{lrtest randslope randint}
\]

Likelihood-ratio test

LR chi2(1) = 291.78

(Assumption: randint nested in randslope)

Prob > chi2 = 0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the parameter space. If this is not true, then the reported test is conservative.

The near-zero significance level favors the model that allows for a random pig-specific regression line over the model that allows only for a pig-specific shift.

Covariance structures

In example 2, we fit a model with the default Independent covariance given in (6). Within any random-effects level specification, we can override this default by specifying an alternative covariance structure via the covariance() option.

Example 3

We generalize (6) to allow \(u_{0j}\) and \(u_{1j}\) to be correlated; that is,

\[
\Sigma = \text{Var} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{01} \\ \sigma_{01} & \sigma_{u1}^2 \end{bmatrix}
\]
. mixed weight week || id: week, covariance(unstructured)
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0:  log likelihood = -868.96185
Iteration 1:  log likelihood = -868.96185
Computing standard errors:
Mixed-effects ML regression  Number of obs = 432
     Group variable: id  Number of groups = 48
                Obs per group: min = 9
                                      avg = 9.0
                                      max = 9
     Wald chi2(1) = 4649.17
Log likelihood = -868.96185  Prob > chi2 = 0.0000

weight
Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
       week  6.209896   .0910745   68.18   0.000      6.031393    6.388399
      _cons  19.35561   .3996387   48.43   0.000     18.57234    20.13889

Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
       id: Unstructured          |                  
               var(week)  .3715251   .0812958     .2419532    .570486
               var(_cons)  6.823363   1.566194     4.351297    10.69986
               cov(week,_cons)  -.0984378   .2545767    -.5973991    .4005234
               var(Residual)  1.596829   .1231980     1.372735    1.857505

LR test vs. linear regression:  chib(3) =  764.58  Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

But we do not find the correlation to be at all significant.

. lrtest . randslope
Likelihood-ratio test  LR chi2(1) =  0.15
(Assumption: randslope nested in .)  Prob > chi2 =  0.6959

Instead, we could have also specified covariance(identity), restricting $u_{0j}$ and $u_{1j}$ to not only be independent but also to have common variance, or we could have specified covariance(exchangeable), which imposes a common variance but allows for a nonzero correlation.

### Likelihood versus restricted likelihood

Thus far, all our examples have used ML to estimate variance components. We could have just as easily asked for REML estimates. Refitting the model in example 2 by REML, we get
. mixed weight week || id: week, reml
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0: log restricted-likelihood = -870.51473
Iteration 1: log restricted-likelihood = -870.51473
Computing standard errors:
Mixed-effects REML regression
Group variable: id
Number of obs = 432
Number of groups = 48
Obs per group: min = 9
avg = 9.0
max = 9
Log restricted-likelihood = -870.51473
Wald chi2(1) = 4592.10
Prob > chi2 = 0.0000

| weight | Coef. | Std. Err. | z | P>|z| | [95% Conf. Interval] |
|--------|-------|-----------|---|-----|----------------------|
| week   | 6.209896 | .0916387  | 67.77 | 0.000 | 6.030287 to 6.389504 |
| _cons  | 19.35561 | .4021144  | 48.13 | 0.000 | 18.56748 to 20.14374 |

Random-effects Parameters | Estimate | Std. Err. | [95% Conf. Interval] |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>id: Independent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(week)</td>
<td>.3764405</td>
<td>.0827027</td>
<td>.2447317 to .5790317</td>
</tr>
<tr>
<td>var(_cons)</td>
<td>6.917604</td>
<td>1.593247</td>
<td>4.404624 to 10.86432</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>1.598784</td>
<td>.1234011</td>
<td>1.374328 to 1.859898</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: chi2(2) = 765.92 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

Although ML estimators are based on the usual likelihood theory, the idea behind REML is to transform the response into a set of linear contrasts whose distribution is free of the fixed effects $\beta$. The restricted likelihood is then formed by considering the distribution of the linear contrasts. Not only does this make the maximization problem free of $\beta$, it also incorporates the degrees of freedom used to estimate $\beta$ into the estimation of the variance components. This follows because, by necessity, the rank of the linear contrasts must be less than the number of observations.

As a simple example, consider a constant-only regression where $y_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. The ML estimate of $\sigma^2$ can be derived theoretically as the $n$-divided sample variance. The REML estimate can be derived by considering the first $n-1$ error contrasts, $y_i - \bar{y}$, whose joint distribution is free of $\mu$. Applying maximum likelihood to this distribution results in an estimate of $\sigma^2$, that is, the $(n-1)$-divided sample variance, which is unbiased for $\sigma^2$.

The unbiasedness property of REML extends to all mixed models when the data are balanced, and thus REML would seem the clear choice in balanced-data problems, although in large samples the difference between ML and REML is negligible. One disadvantage of REML is that likelihood-ratio (LR) tests based on REML are inappropriate for comparing models with different fixed-effects specifications. ML is appropriate for such LR tests and has the advantage of being easy to explain and being the method of choice for other estimators.

Another factor to consider is that ML estimation under *mixed* is more feature-rich, allowing for weighted estimation and robust variance–covariance matrices, features not supported under REML. In the end, which method to use should be based both on your needs and on personal taste.
Examining the REML output, we find that the estimates of the variance components are slightly larger than the ML estimates. This is typical, because ML estimates, which do not incorporate the degrees of freedom used to estimate the fixed effects, tend to be biased downward.

Three-level models

The clustered-data representation of the mixed model given in (2) can be extended to two nested levels of clustering, creating a three-level model once the observations are considered. Formally,

$$y_{jk} = X_{jk} \beta + Z_{jk}^{(3)} u_k^{(3)} + Z_{jk}^{(2)} u_{jk}^{(2)} + \epsilon_{jk}$$

for $i = 1, \ldots, n_{jk}$ first-level observations nested within $j = 1, \ldots, M_k$ second-level groups, which are nested within $k = 1, \ldots, M$ third-level groups. Group $j, k$ consists of $n_{jk}$ observations, so $y_{jk}$, $X_{jk}$, and $\epsilon_{jk}$ each have row dimension $n_{jk}$. $Z_{jk}^{(3)}$ is the $n_{jk} \times q_3$ design matrix for the third-level random effects $u_k^{(3)}$, and $Z_{jk}^{(2)}$ is the $n_{jk} \times q_2$ design matrix for the second-level random effects $u_{jk}^{(2)}$. Furthermore, assume that

$$u_k^{(3)} \sim N(0, \Sigma_3); \quad u_{jk}^{(2)} \sim N(0, \Sigma_2); \quad \epsilon_{jk} \sim N(0, \sigma^2_{\epsilon} I)$$

and that $u_k^{(3)}$, $u_{jk}^{(2)}$, and $\epsilon_{jk}$ are independent.

Fitting a three-level model requires you to specify two random-effects equations: one for level three and then one for level two. The variable list for the first equation represents $Z_{jk}^{(3)}$ and for the second equation represents $Z_{jk}^{(2)}$; that is, you specify the levels top to bottom in `mixed`.

Example 4

Baltagi, Song, and Jung (2001) estimate a Cobb–Douglas production function examining the productivity of public capital in each state’s private output. Originally provided by Munnell (1990), the data were recorded over 1970–1986 for 48 states grouped into nine regions.
. use http://www.stata-press.com/data/r13/productivity
(Public Capital Productivity)
. describe
Contains data from http://www.stata-press.com/data/r13/productivity.dta
obs: 816 Public Capital Productivity
vars: 11 29 Mar 2013 10:57
size: 29,376 (_dta has notes)

<table>
<thead>
<tr>
<th>variable name</th>
<th>storage type</th>
<th>display format</th>
<th>value label</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>byte</td>
<td>%9.0g</td>
<td>states 1-48</td>
</tr>
<tr>
<td>region</td>
<td>byte</td>
<td>%9.0g</td>
<td>regions 1-9</td>
</tr>
<tr>
<td>year</td>
<td>int</td>
<td>%9.0g</td>
<td>years 1970-1986</td>
</tr>
<tr>
<td>public</td>
<td>float</td>
<td>%9.0g</td>
<td>public capital stock</td>
</tr>
<tr>
<td>hwy</td>
<td>float</td>
<td>%9.0g</td>
<td>log(highway component of public)</td>
</tr>
<tr>
<td>water</td>
<td>float</td>
<td>%9.0g</td>
<td>log(water component of public)</td>
</tr>
<tr>
<td>other</td>
<td>float</td>
<td>%9.0g</td>
<td>log(bldg/other component of public)</td>
</tr>
<tr>
<td>private</td>
<td>float</td>
<td>%9.0g</td>
<td>log(private capital stock)</td>
</tr>
<tr>
<td>gsp</td>
<td>float</td>
<td>%9.0g</td>
<td>log(gross state product)</td>
</tr>
<tr>
<td>emp</td>
<td>float</td>
<td>%9.0g</td>
<td>log(non-agriculture payrolls)</td>
</tr>
<tr>
<td>unemp</td>
<td>float</td>
<td>%9.0g</td>
<td>state unemployment rate</td>
</tr>
</tbody>
</table>

Sorted by:

Because the states are nested within regions, we fit a three-level mixed model with random intercepts at both the region and the state-within-region levels. That is, we use (7) with both $Z_{jk}^{(3)}$ and $Z_{jk}^{(2)}$ set to the $n_{jk} 	imes 1$ column of ones, and $\Sigma_3 = \sigma_3^2$ and $\Sigma_2 = \sigma_2^2$ are both scalars.

. mixed gsp private emp hwy water other unemp || region: || state:
(output omitted)

Mixed-effects ML regression

<table>
<thead>
<tr>
<th>Group Variable</th>
<th>No. of Groups</th>
<th>Observations per Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum</td>
<td>Average</td>
</tr>
<tr>
<td>region</td>
<td>9</td>
<td>51</td>
</tr>
<tr>
<td>state</td>
<td>48</td>
<td>17</td>
</tr>
</tbody>
</table>

Wald chi2(6) = 18829.06
Prob > chi2 = 0.0000

Log likelihood = 1430.5017

<p>| gsp   | Coef. | Std. Err. | z     | P&gt;|z|  | [95% Conf. Interval] |
|-------|-------|-----------|-------|------|----------------------|
| private | .2671484 | .0212591 | 12.57 | 0.000 | .2254814 .3088154 |
| emp    | .754072  | .0261868  | 28.80 | 0.000 | .7027468 .8053973 |
| hwy    | .0709767 | .023041   | 3.08  | 0.002 | .0258172 .1161363 |
| water  | .0761187 | .0139248  | 5.47  | 0.000 | .0488266 .1034109 |
| other  | -.0999955 | .0169366 | -5.90 | 0.000 | -.1331906 -.0689004 |
| unemp  | -.0058983 | .0009031 | -6.53 | 0.000 | -.0076684 -.0041282 |
| _cons  | 2.128823 | .1543854  | 13.79 | 0.000 | 1.826233 2.431413 |</p>
<table>
<thead>
<tr>
<th>Random-effects Parameters</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>region: Identity var(_cons)</td>
<td>.0014506</td>
<td>.0012995</td>
<td>.0002506 .0083957</td>
</tr>
<tr>
<td>state: Identity var(_cons)</td>
<td>.0062757</td>
<td>.0014871</td>
<td>.0039442 .0099855</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>.0013461</td>
<td>.0000689</td>
<td>.0012176 .0014882</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: $\text{chi}^2(2) = 1154.73$  Prob $> \text{chi}^2 = 0.0000$

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

Notes:
1. Our model now has two random-effects equations, separated by ||. The first is a random intercept (constant only) at the region level (level three), and the second is a random intercept at the state level (level two). The order in which these are specified (from left to right) is significant—mixed assumes that state is nested within region.

2. The information on groups is now displayed as a table, with one row for each grouping. You can suppress this table with the nogroup or the noheader option, which will suppress the rest of the header, as well.

3. The variance-component estimates are now organized and labeled according to level.

   After adjusting for the nested-level error structure, we find that the highway and water components of public capital had significant positive effects on private output, whereas the other public buildings component had a negative effect.

Technical note

In the previous example, the states are coded 1–48 and are nested within nine regions. mixed treated the states as nested within regions, regardless of whether the codes for each state were unique between regions. That is, even if codes for states were duplicated between regions, mixed would have enforced the nesting and produced the same results.

The group information at the top of the mixed output and that produced by the postestimation command estat group (see [ME] mixed postestimation) take the nesting into account. The statistics are thus not necessarily what you would get if you instead tabulated each group variable individually.

Model (7) extends in a straightforward manner to more than three levels, as does the specification of such models in mixed.

Blocked-diagonal covariance structures

Covariance matrices of random effects within an equation can be modeled either as a multiple of the identity matrix, as diagonal (that is, Independent), as exchangeable, or as general symmetric (Unstructured). These may also be combined to produce more complex block-diagonal covariance structures, effectively placing constraints on the variance components.
Example 5

Returning to our productivity data, we now add random coefficients on hwy and unemp at the region level. This only slightly changes the estimates of the fixed effects, so we focus our attention on the variance components:

```
mixed gsp private emp hwy water other unemp || region: hwy unemp || state:,
> nolog nogroup nofetable
```

Mixed-effects ML regression

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>region: Independent</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(hwy)</td>
<td>0.0000209</td>
<td>0.0001103</td>
<td>6.71e-10 · 6.50695</td>
</tr>
<tr>
<td>var(unemp)</td>
<td>0.0000238</td>
<td>0.0000135</td>
<td>7.84e-06 · 0.000722</td>
</tr>
<tr>
<td>var(_cons)</td>
<td>0.0030349</td>
<td>0.0086684</td>
<td>0.000112 · 0.8191296</td>
</tr>
<tr>
<td>state: Identity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(_cons)</td>
<td>0.0063658</td>
<td>0.0015611</td>
<td>0.0039365 · 0.0102943</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>0.0012469</td>
<td>0.0000643</td>
<td>0.001127 · 0.0013795</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: chi2(4) = 1189.08 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

This model is the same as that fit in example 4 except that $Z_{jk}^{(3)}$ is now the $n_{jk} \times 3$ matrix with columns determined by the values of hwy, unemp, and an intercept term (one), in that order, and (because we used the default Independent structure) $\Sigma_3$ is

$$
\Sigma_3 = \begin{pmatrix}
\sigma_a^2 & 0 & 0 \\
0 & \sigma_b^2 & 0 \\
0 & 0 & \sigma_c^2 
\end{pmatrix}
$$

The random-effects specification at the state level remains unchanged; that is, $\Sigma_2$ is still treated as the scalar variance of the random intercepts at the state level.

An LR test comparing this model with that from example 4 favors the inclusion of the two random coefficients, a fact we leave to the interested reader to verify.

The estimated variance components, upon examination, reveal that the variances of the random coefficients on hwy and unemp could be treated as equal. That is,

$$
\Sigma_3 = \begin{pmatrix}
\sigma_a^2 & 0 & 0 \\
0 & \sigma_a^2 & 0 \\
0 & 0 & \sigma_c^2 
\end{pmatrix}
$$

looks plausible. We can impose this equality constraint by treating $\Sigma_3$ as block diagonal: the first block is a $2 \times 2$ multiple of the identity matrix, that is, $\sigma_a^2 I_2$; the second is a scalar, equivalently, a $1 \times 1$ multiple of the identity.
We construct block-diagonal covariances by repeating level specifications:

```
.mixed gsp private emp hwy water other unemp || region: hwy unemp,>
cov(identity) || region: || state:, nolog nogroup nofetable
```

Mixed-effects ML regression

<table>
<thead>
<tr>
<th>Number of obs</th>
<th>816</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald chi2(6)</td>
<td>17136.65</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Random-effects Parameters  | Estimate  | Std. Err.  | [95% Conf. Interval] |
---------------------------|-----------|------------|----------------------|
region: Identity          |           |            |                      |
  var(hwy unemp)           | 0.0000238 | 0.0000134  | 7.89e-06             |
                          |           |            | 0.000719             |
region: Identity          |           |            |                      |
  var(_cons)               | 0.0028191 | 0.0030429  | 0.0003399            |
                          |           |            | 0.023383             |
state: Identity           |           |            |                      |
  var(_cons)               | 0.006358  | 0.0015309  | 0.0039661            |
                          |           |            | 0.0101925            |
  var(Residual)            | 0.0012469 | 0.0000643  | 0.001127             |
                          |           |            | 0.0013795            |

LR test vs. linear regression:  chi2(3) = 1189.08  Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

We specified two equations for the `region` level: the first for the random coefficients on `hwy` and `unemp` with covariance set to `Identity` and the second for the random intercept `cons`, whose covariance defaults to `Identity` because it is of dimension 1. `mixed` labeled the estimate of $\sigma^2_a$ as `var(hwy unemp)` to designate that it is common to the random coefficients on both `hwy` and `unemp`.

An LR test shows that the constrained model fits equally well.

```
.lrtest prodrc
```

Likelihood-ratio test  LR chi2(1) = 0.00  Prob > chi2 = 0.9784
(Assumption: . nested in prodrc)  Prob > chi2 = 0.9784

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the parameter space. If this is not true, then the reported test is conservative.

Because the null hypothesis for this test is one of equality ($H_0: \sigma^2_a = \sigma^2_b$), it is not on the boundary of the parameter space. As such, we can take the reported significance as precise rather than a conservative estimate.

You can repeat level specifications as often as you like, defining successive blocks of a block-diagonal covariance matrix. However, repeated-level equations must be listed consecutively; otherwise, `mixed` will give an error.

## Technical note

In the previous estimation output, there was no constant term included in the first `region` equation, even though we did not use the `noconstant` option. When you specify repeated-level equations, `mixed` knows not to put constant terms in each equation because such a model would be unidentified. By default, it places the constant in the last repeated-level equation, but you can use `noconstant` creatively to override this.

Linear mixed-effects models can also be fit using `meglm` with the default `gaussian` family. `meglm` provides two more covariance structures through which you can impose constraints on variance components; see [ME] `meglm` for details.
Heteroskedastic random effects

Blocked-diagonal covariance structures and repeated-level specifications of random effects can also be used to model heteroskedasticity among random effects at a given level.

Example 6

Following Rabe-Hesketh and Skrondal (2012, sec. 7.2), we analyze data from Asian children in a British community who were weighed up to four times, roughly between the ages of 6 weeks and 27 months. The dataset is a random sample of data previously analyzed by Goldstein (1986) and Prosser, Rasbash, and Goldstein (1991).

```
. use http://www.stata-press.com/data/r13/childweight
(Weight data on Asian children)
. describe
Contains data from http://www.stata-press.com/data/r13/childweight.dta
    obs: 198                Weight data on Asian children
    vars: 5                  23 May 2013 15:12
    size: 3,168          (_dta has notes)

+-----------------+----------------------------------+
| variable name   | storage | display | value | variable label       |
|-----------------+----------|---------|-------|----------------------|
| id              | int      | %8.0g  |       | child identifier     |
| age             | float    | %8.0g  |       | age in years         |
| weight          | float    | %8.0g  |       | weight in Kg         |
| brthwt          | int      | %8.0g  |       | Birth weight in g    |
| girl            | float    | %9.0g  | bg    | gender               |
+-----------------+----------|---------|-------|----------------------+

Sorted by:  id  age
```

```
. graph twoway (line weight age, connect(ascending)), by(girl)
> xtitle(Age in years) ytitle(Weight in kg)
```

Ignoring gender effects for the moment, we begin with the following model for the $i$th measurement on the $j$th child:

\[
weight_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 age_{ij}^2 + u_{j0} + u_{j1} age_{ij} + \epsilon_{ij}
\]
This models overall mean growth as quadratic in age and allows for two child-specific random effects: a random intercept $u_{j0}$, which represents each child’s vertical shift from the overall mean ($\beta_0$), and a random age slope $u_{j1}$, which represents each child’s deviation in linear growth rate from the overall mean linear growth rate ($\beta_1$). For simplicity, we do not consider child-specific changes in the quadratic component of growth.

. mixed weight age c.age#c.age || id: age, nolog

Mixed-effects ML regression

|               | Coef.    | Std. Err. | z     | P>|z|    | [95% Conf. Interval] |
|---------------|----------|-----------|-------|--------|---------------------|
| age           | 7.693701 | .2381076  | 32.31 | 0.000  | 7.227019 8.160384   |
| c.age#c.age   | -1.654542| .0874987  | -18.91| 0.000  | -1.826037 -1.483048 |
| _cons         | 3.497628 | .1416914  | 24.68 | 0.000  | 3.219918 3.775338   |

Random-effects Parameters

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
</table>
| id: Independent
  var(age)      | .2987207 | .0827569  | .1735603 .5141388   |
  var(_cons)    | .5023857 | .141263   | .2895294 .8717297   |
  var(Residual) | .3092897 | .0474887  | .2289133 .417888    |

LR test vs. linear regression: chi2(2) = 114.70 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

Because there is no reason to believe that the random effects are uncorrelated, it is always a good idea to first fit a model with the covariance(unstructured) option. We do not include the output for such a model because for these data the correlation between random effects is not significant; however, we did check this before reverting to mixed’s default Independent structure.

Next we introduce gender effects into the fixed portion of the model by including a main gender effect and a gender–age interaction for overall mean growth:
The main gender effect is significant at the 5% level, but the gender–age interaction is not:

```
. test 0.girl#c.age = 1.girl#c.age
( 1) [weight]0b.girl#c.age - [weight]1.girl#c.age = 0
    chi2( 1) =    1.66
    Prob > chi2 =   0.1978
```

On average, boys are heavier than girls, but their average linear growth rates are not significantly different.

In the above model, we introduced a gender effect on average growth, but we still assumed that the variability in child-specific deviations from this average was the same for boys and girls. To check this assumption, we introduce gender into the random component of the model. Because support for factor-variable notation is limited in specifications of random effects (see Crossed-effects models below), we need to generate the interactions ourselves.
. gen boy = !girl
. gen boyXage = boy*age
. gen girlXage = girl*age

. mixed weight i.girl i.girl#c.age c.age#c.age || id: boy boyXage, noconstant
> || id: girl girlXage, noconstant nolog nofetable

Mixed-effects ML regression

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>id: Independent</td>
<td>var(boy)</td>
<td>.3161091</td>
<td>.1557911</td>
</tr>
<tr>
<td></td>
<td>var(boyXage)</td>
<td>.4734482</td>
<td>.1574626</td>
</tr>
<tr>
<td>id: Independent</td>
<td>var(girl)</td>
<td>.5798676</td>
<td>.1959725</td>
</tr>
<tr>
<td></td>
<td>var(girlXage)</td>
<td>.0664634</td>
<td>.0553274</td>
</tr>
<tr>
<td></td>
<td>var(Residual)</td>
<td>.3078826</td>
<td>.046484</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: chi2(4) = 112.86  Prob > chi2 = 0.0000

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

In the above, we suppress displaying the fixed portion of the model (the nofetable option) because it does not differ much from that of the previous model.

Our previous model had the random-effects specification

|| id: age

which we have replaced with the dual repeated-level specification

|| id: boy boyXage, noconstant || id: girl girlXage, noconstant

The former models a random intercept and random slope on age, and does so treating all children as a random sample from one population. The latter also specifies a random intercept and random slope on age, but allows for the variability of the random intercepts and slopes to differ between boys and girls. In other words, it allows for heteroskedasticity in random effects due to gender. We use the noconstant option so that we can separate the overall random intercept (automatically provided by the former syntax) into one specific to boys and one specific to girls.

There seems to be a large gender effect in the variability of linear growth rates. We can compare both models with an LR test, recalling that we stored the previous estimation results under the name homoskedastic:

. lrtest homoskedastic heteroskedastic

Likelihood-ratio test

<table>
<thead>
<tr>
<th></th>
<th>LR ch2(2)</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Assumption: homoskedastic nested in heteroskedastic)</td>
<td>8.47</td>
<td>0.0145</td>
</tr>
</tbody>
</table>

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the parameter space. If this is not true, then the reported test is conservative.
Because the null hypothesis here is one of equality of variances and not that variances are 0, the above does not test on the boundary; thus we can treat the significance level as precise and not conservative. Either way, the results favor the new model with heteroskedastic random effects.

**Heteroskedastic residual errors**

Up to this point, we have assumed that the level-one residual errors—the $\epsilon$’s in the stated models—have been i.i.d. Gaussian with variance $\sigma_\epsilon^2$. This is demonstrated in `mixed` output in the random-effects table, where up until now we have estimated a single residual-error variance, labeled as $\text{var(Residual)}$.

To relax the assumptions of homoskedasticity or independence of residual errors, use the `residuals()` option.

> **Example 7**

West, Welch, and Galecki (2007, chap. 7) analyze data studying the effect of ceramic dental veneer placement on gingival (gum) health. Data on 55 teeth located in the maxillary arches of 12 patients were considered.

```
. use http://www.stata-press.com/data/r13/veneer, clear
   (Dental veneer data)
. describe
Contains data from http://www.stata-press.com/data/r13/veneer.dta
   obs: 110   Dental veneer data
   vars:  7   24 May 2013 12:11
   size: 1,100 (_dta has notes)

     variable name storage display label
    --------- ------- ----------- ---------------------------
     patient   byte   %8.0g    Patient ID
     tooth     byte   %8.0g    Tooth number with patient
     gcf       byte   %8.0g    Gingival crevicular fluid (GCF)
     age       byte   %8.0g    Patient age
     base_gcf  byte   %8.0g    Baseline GCF
     cda       float   %9.0g   Average contour difference after veneer placement
     followup  byte   %9.0g   t Follow-up time: 3 or 6 months
```

Sort by:

Veneers were placed to match the original contour of the tooth as closely as possible, and researchers were interested in how contour differences (variable `cda`) impacted gingival health. Gingival health was measured as the amount of gingival crevical fluid (GCF) at each tooth, measured at baseline (variable `base_gcf`) and at two posttreatment follow-ups at 3 and 6 months. The variable `gcf` records GCF at follow-up, and the variable `followup` records the follow-up time.

Because two measurements were taken for each tooth and there exist multiple teeth per patient, we fit a three-level model with the following random effects: a random intercept and random slope on follow-up time at the patient level, and a random intercept at the tooth level. For the $i$th measurement of the $j$th tooth from the $k$th patient, we have

\[
gcf_{ijk} = \beta_0 + \beta_1\text{followup}_{ijk} + \beta_2\text{base_gcf}_{ijk} + \beta_3\text{cda}_{ijk} + \beta_4\text{age}_{ijk} + u_{0k} + u_{1k}\text{followup}_{ijk} + v_{0jk} + \epsilon_{ijk}
\]
which we can fit using `mixed`:

```
.mixed gcf followup base_gcf cda age || patient: followup, cov(un) || tooth:, > reml nolog
```

Mixed-effects REML regression

<table>
<thead>
<tr>
<th>Group Variable</th>
<th>No. of Groups</th>
<th>Observations per Group</th>
<th>Wald chi2(4)</th>
<th>Log restricted-likelihood</th>
<th>Prob &gt; chi2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-420.92761</td>
<td>0.1128</td>
</tr>
</tbody>
</table>

| gcf            | Coef.        | Std. Err.  | z    | P>|z|    | [95% Conf. Interval] |
|----------------|--------------|------------|------|---------|----------------------|
| followup       | .3009815     | 1.936863   | 0.16 | 0.877   | -3.4952, 4.097163    |
| base_gcf       | -.0183127    | .1433094   | -0.13| 0.898   | -.299194, .2625685  |
| cda            | -.329303     | .5292525   | -0.62| 0.534   | -1.366619, .7080128 |
| age            | -.5773932    | .2139656   | -2.70| 0.007   | -.9967582, -.1580283|
| _cons          | 45.73862     | 12.55497   | 3.64 | 0.000   | 21.13133, 70.34591  |

<table>
<thead>
<tr>
<th>Random-effects Parameters</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient: Unstructured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(followup)</td>
<td>41.88772</td>
<td>18.79997</td>
<td>17.38009, 100.9535</td>
</tr>
<tr>
<td>var(_cons)</td>
<td>524.9851</td>
<td>253.0205</td>
<td>204.0287, 1350.175</td>
</tr>
<tr>
<td>cov(followup,_cons)</td>
<td>-140.4229</td>
<td>66.57623</td>
<td>-270.9099, -9.935908</td>
</tr>
<tr>
<td>tooth: Identity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(_cons)</td>
<td>47.45738</td>
<td>16.63034</td>
<td>23.8792, 94.3165</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>48.86704</td>
<td>10.50523</td>
<td>32.06479, 74.47382</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: chi2(4) = 91.12 Prob > chi2 = 0.0000

We used REML estimation for no other reason than variety.

Among the other features of the model fit, we note that the residual variance $\sigma^2$ was estimated as 48.87 and that our model assumed that the residuals were independent with constant variance (homoskedastic). Because it may be the case that the precision of gcf measurements could change over time, we modify the above to estimate two distinct error variances: one for the 3-month follow-up and one for the 6-month follow-up.

To fit this model, we add the `residuals(independent, by(followup))` option, which maintains independence of residual errors but allows for heteroskedasticity with respect to follow-up time.
. mixed gcf followup base_gcf cda age || patient: followup, cov(un) || tooth:, 
> residuals(independent, by(followup)) reml nolog

Mixed-effects REML regression

<table>
<thead>
<tr>
<th>Group Variable</th>
<th>No. of Groups</th>
<th>Observations per Group</th>
<th>Number of obs = 110</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum</td>
<td>Average</td>
<td>Maximum</td>
</tr>
<tr>
<td>patient</td>
<td>12</td>
<td>2</td>
<td>9.2</td>
</tr>
<tr>
<td>tooth</td>
<td>55</td>
<td>2</td>
<td>2.0</td>
</tr>
</tbody>
</table>

Wald chi2(4) = 7.51
Log restricted-likelihood = -420.4576
Prob > chi2 = 0.1113

| gcf          | Coef.  | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|--------------|--------|-----------|-------|-------|----------------------|
| followup     | .2703944 | 1.933096 | 0.14  | 0.889 | -3.518405 4.059193   |
| base_gcf     | 0.0062144 | .1419121 | 0.04  | 0.965 | -.2719283 .284357    |
| cda          | -.2947235 | .5245126 | -0.56 | 0.574 | -1.322749 .7333023   |
| age          | -.5743755 | .2142249 | -2.68 | 0.007 | -1.9942487 -.1545024 |
| _cons        | 45.15089 | 12.51452 | 3.61  | 0.000 | 20.62288 69.6789     |

Random-effects Parameters

<table>
<thead>
<tr>
<th>Random-effects Parameters</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>patient: Unstructured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(followup)</td>
<td>41.75169</td>
<td>18.72989</td>
<td>17.33099 100.583</td>
</tr>
<tr>
<td>var(_cons)</td>
<td>515.2018</td>
<td>251.9661</td>
<td>197.5542 1343.596</td>
</tr>
<tr>
<td>cov(followup,_cons)</td>
<td>-139.0496</td>
<td>66.27806</td>
<td>-268.9522 -9.14694</td>
</tr>
<tr>
<td>tooth: Identity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(_cons)</td>
<td>47.35914</td>
<td>16.48931</td>
<td>23.93514 93.70693</td>
</tr>
</tbody>
</table>

Residual: Independent, 
by followup

| 3 months: var(e) | 61.36785 | 18.38913 | 34.10946 | 110.4096 |
| 6 months: var(e) | 36.42861 | 14.97501 | 16.27542 | 81.53666 |

LR test vs. linear regression: chi2(5) = 92.06 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

Comparison of both models via an LR test reveals the difference in residual variances to be not significant, something we leave to you to verify as an exercise.

The default residual-variance structure is independent, and when specified without by() is equivalent to the default behavior of mixed: estimating one overall residual standard variance for the entire model.

Other residual-error structures

Besides the default independent residual-error structure, mixed supports four other structures that allow for correlation between residual errors within the lowest-level (smallest or level two) groups. For purposes of notation, in what follows we assume a two-level model, with the obvious extension to higher-level models.

The exchangeable structure assumes one overall variance and one common pairwise covariance; that is,
Var(ε_j) = Var
\[
\begin{bmatrix}
ε_{j1} \\
ε_{j2} \\
\vdots \\
ε_{jn}
\end{bmatrix}
= \begin{bmatrix}
σ_ε^2 & σ_1 & \cdots & σ_1 \\
σ_1 & σ_ε^2 & \cdots & σ_1 \\
\vdots & \vdots & \ddots & \vdots \\
σ_1 & σ_1 & \cdots & σ_ε^2
\end{bmatrix}
\]

By default, mixed will report estimates of the two parameters as estimates of the common variance \(σ_ε^2\) and of the covariance \(σ_1\). If the stddeviations option is specified, you obtain estimates of \(σ_ε\) and the pairwise correlation. When the by(varname) option is also specified, these two parameters are estimated for each level varname.

The ar p structure assumes that the errors have an AR structure of order p. That is,
\[
ε_{ij} = φ_1ε_{i-1,j} + \cdots + φ_pε_{i-p,j} + u_{ij}
\]
where \(u_{ij}\) are i.i.d. Gaussian with mean 0 and variance \(σ_u^2\). mixed reports estimates of \(φ_1, \ldots, φ_p\) and the overall error variance \(σ_ε^2\), which can be derived from the above expression. The t(varname) option is required, where varname is a time variable used to order the observations within lowest-level groups and to determine any gaps between observations. When the by(varname) option is also specified, the set of \(p + 1\) parameters is estimated for each level of varname. If \(p = 1\), then the estimate of \(φ_1\) is reported as rho, because in this case it represents the correlation between successive error terms.

The ma q structure assumes that the errors are an MA process of order q. That is,
\[
ε_{ij} = u_{ij} + θ_1u_{i-1,j} + \cdots + θ_qu_{i-q,j}
\]
where \(u_{ij}\) are i.i.d. Gaussian with mean 0 and variance \(σ_u^2\). mixed reports estimates of \(θ_1, \ldots, θ_q\) and the overall error variance \(σ_ε^2\), which can be derived from the above expression. The t(varname) option is required, where varname is a time variable used to order the observations within lowest-level groups and to determine any gaps between observations. When the by(varname) option is also specified, the set of \(q + 1\) parameters is estimated for each level of varname.

The unstructured structure is the most general and estimates unique variances and unique pairwise covariances for all residuals within the lowest-level grouping. Because the data may be unbalanced and the ordering of the observations is arbitrary, the t(varname) option is required, where varname is an identification variable that matches error terms in different groups. If varname has \(n\) distinct levels, then \(n(n + 1)/2\) parameters are estimated. Not all \(n\) levels need to be observed within each group, but duplicated levels of varname within a given group are not allowed because they would cause a singularity in the estimated error-variance matrix for that group. When the by(varname) option is also specified, the set of \(n(n + 1)/2\) parameters is estimated for each level of varname.

The banded q structure is a special case of unstructured that confines estimation to within the first q off-diagonal elements of the residual variance–covariance matrix and sets the covariances outside this band to 0. As is the case with unstructured, the t(varname) option is required, where varname is an identification variable that matches error terms in different groups. However, with banded variance structures, the ordering of the values in varname is significant because it determines which covariances are to be estimated and which are to be set to 0. For example, if varname has \(n = 5\) distinct values \(t = 1, 2, 3, 4, 5\), then a banded variance–covariance structure of order \(q = 2\) would estimate the following:
\[
\text{Var}(\epsilon_j) = \text{Var}
\begin{bmatrix}
\epsilon_{1j} \\
\epsilon_{2j} \\
\epsilon_{3j} \\
\epsilon_{4j} \\
\epsilon_{5j}
\end{bmatrix} = \begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \sigma_{13} & 0 & 0 \\
\sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & 0 \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\
0 & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{45} \\
0 & 0 & \sigma_{35} & \sigma_{45} & \sigma_5^2
\end{bmatrix}
\]

In other words, you would have an unstructured variance matrix that constrains \(\sigma_{14} = \sigma_{15} = \sigma_{25} = 0\). If \(\text{varname}\) has \(n\) distinct levels, then \((q + 1)(2n - q)/2\) parameters are estimated. Not all \(n\) levels need to be observed within each group, but duplicated levels of \(\text{varname}\) within a given group are not allowed because they would cause a singularity in the estimated error-variance matrix for that group. When the \(\text{by} (\text{varname})\) option is also specified, the set of parameters is estimated for each level of \(\text{varname}\). If \(q\) is left unspecified, then \(\text{banded}\) is equivalent to \(\text{unstructured}\); that is, all variances and covariances are estimated. When \(q = 0\), \(\text{Var}(\epsilon_j)\) is treated as diagonal and can thus be used to model uncorrelated yet heteroskedastic residual errors.

The \(\text{toeplitz} q\) structure assumes that the residual errors are homoskedastic and that the correlation between two errors is determined by the time lag between the two. That is, \(\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2\) and

\[
\text{Corr}(\epsilon_{ij}, \epsilon_{i+k,j}) = \rho_k
\]

If the lag \(k\) is less than or equal to \(q\), then the pairwise correlation \(\rho_k\) is estimated; if the lag is greater than \(q\), then \(\rho_k\) is assumed to be 0. If \(q\) is left unspecified, then \(\rho_k\) is estimated for each observed lag \(k\). The \(\text{t} (\text{varname})\) option is required, where \(\text{varname}\) is a time variable \(t\) used to determine the lags between pairs of residual errors. As such, \(t()\) must be integer-valued. \(q + 1\) parameters are estimated: one overall variance \(\sigma_\epsilon^2\) and \(q\) correlations. When the \(\text{by} (\text{varname})\) option is also specified, the set of \(q + 1\) parameters is estimated for each level of \(\text{varname}\).

The \(\text{exponential}\) structure is a generalization of the \(\text{AR}\) structure that allows for noninteger and irregularly spaced time lags. That is, \(\text{Var}(\epsilon_{ij}) = \sigma_\epsilon^2\) and

\[
\text{Corr}(\epsilon_{ij}, \epsilon_{kj}) = \rho^{|i-k|}
\]

for \(0 \leq \rho \leq 1\), with \(i\) and \(k\) not required to be integers. The \(\text{t} (\text{varname})\) option is required, where \(\text{varname}\) is a time variable used to determine \(i\) and \(k\) for each residual-error pair. \(t()\) is real-valued. \(\text{mixed}\) reports estimates of \(\sigma_\epsilon^2\) and \(\rho\). When the \(\text{by} (\text{varname})\) option is also specified, these two parameters are estimated for each level of \(\text{varname}\).

Example 8

Pinheiro and Bates (2000, chap. 5) analyze data from a study of the estrus cycles of mares. Originally analyzed in Pierson and Ginther (1987), the data record the number of ovarian follicles larger than 10mm, daily over a period ranging from three days before ovulation to three days after the subsequent ovulation.
. use http://www.stata-press.com/data/r13/ovary
(0varian follicles in mares)
. describe
Contains data from http://www.stata-press.com/data/r13/ovary.dta
    obs: 308 Ovarian follicles in mares
    vars: 6 20 May 2013 13:49
    size: 5,544 (_dta has notes)

storage  display  value
variable name type format label variable label

mare     byte %9.0g  mare ID
stime    float %9.0g  Scaled time
follicles byte %9.0g  Number of ovarian follicles > 10
               mm in diameter
sin1     float %9.0g  sine(2*π*stime)
cos1     float %9.0g  cosine(2*π*stime)
time     float %9.0g  time order within mare

Sorted by:  mare  stime

The stime variable is time that has been scaled so that ovulation occurs at scaled times 0 and 1, and the time variable records the time ordering within mares. Because graphical evidence suggests a periodic behavior, the analysis includes the sin1 and cos1 variables, which are sine and cosine transformations of scaled time, respectively.

We consider the following model for the $i$th measurement on the $j$th mare:

$$
follicles_{ij} = \beta_0 + \beta_1 \sin1_{ij} + \beta_2 \cos1_{ij} + u_j + \epsilon_{ij}
$$

The above model incorporates the cyclical nature of the data as affecting the overall average number of follicles and includes mare-specific random effects $u_j$. Because we believe successive measurements within each mare are probably correlated (even after controlling for the periodicity in the average), we also model the within-mare errors as being AR of order 2.

. mixed follicles sin1 cos1 || mare:, residuals(ar 2, t(time)) reml nolog

Mixed-effects REML regression
Number of obs = 308
Number of groups = 11
Obs per group: min = 25
               avg = 28.0
               max = 31

Wald chi2(2) = 34.72
Prob > chi2 = 0.0000

Log restricted-likelihood = -772.59855

| follicles | Coef.  | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|-----------|--------|-----------|-------|-------|----------------------|
| sin1      | -2.899228 | .5110786  | -5.67  | 0.000 | -3.900923 -1.897532   |
| cos1      | -.8652936 | .5432926  | -1.59  | 0.111 | -1.930127 .1985402   |
| _cons     | 12.14455  | .9473631  | 12.82  | 0.000 | 10.28775 14.00135    |
Random-effects Parameters | Estimate | Std. Err. | [95% Conf. Interval]
--- | --- | --- | ---
mare: Identity | var(_cons) | 7.092439 | 4.401937 | 2.101337 | 23.93843
Residual: AR(2) | phi1 | .5386104 | .0624899 | .4161325 | .6610883
phi2 | .144671 | .0632041 | .0207933 | .2685488
var(e) | 14.25104 | 2.435238 | 10.19512 | 19.92054

LR test vs. linear regression: chi2(3) = 251.67 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

We picked an order of 2 as a guess, but we could have used LR tests of competing AR models to determine the optimal order, because models of smaller order are nested within those of larger order.

### Example 9

Fitzmaurice, Laird, and Ware (2011, chap. 7) analyzed data on 37 subjects who participated in an exercise therapy trial.

```stata
use http://www.stata-press.com/data/r13/exercise
describe
Contains data from http://www.stata-press.com/data/r13/exercise.dta
obs: 259  Exercise Therapy Trial
vars: 4  24 Jun 2012 18:35
size: 1,036  (_dta has notes)

 storage  display value
variable name  type  format  label                  variable label
---  ---  ------  -------------
id  byte  %9.0g  Person ID
day  byte  %9.0g  Day of measurement
program  byte  %9.0g  1 = reps increase; 2 = weights increase
strength  byte  %9.0g  Strength measurement
```

Subjects (variable `id`) were placed on either an increased-repetition regimen (`program==1`) or a program that kept the repetitions constant but increased weight (`program==2`). Muscle-strength measurements (variable `strength`) were taken at baseline (`day==0`) and then every two days over the next twelve days.

Following Fitzmaurice, Laird, and Ware (2011, chap. 7), and to demonstrate fitting residual-error structures to data collected at uneven time points, we confine our analysis to those data collected at baseline and at days 4, 6, 8, and 12. We fit a full two-way factorial model of `strength` on `program` and `day`, with an unstructured residual-error covariance matrix over those repeated measurements taken on the same subject:
. keep if inlist(day, 0, 4, 6, 8, 12)
(74 observations deleted)
. mixed strength i.program##i.day || id:,
> noconstant residuals(unstructured, t(day)) nolog

Mixed-effects ML regression
Group variable: id
Number of obs = 173
Number of groups = 37
Obs per group: min = 3
avg = 4.7
max = 5

Log likelihood = -296.58215
Prob > chi2 = 0.0000

| strength | Coef. | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|----------|-------|-----------|-------|-------|----------------------|
| 2.program| 1.360119 | 1.003549 | 1.36  | 0.175 | -0.6068016 - 3.32704 |
| day      |       |           |       |       |                      |
| 4        | 1.125 | .3322583 | 3.39  | 0.001 | .4737858 1.776214   |
| 6        | 1.360127 | .3766894 | 3.61  | 0.000 | .6218298 2.098425  |
| 8        | 1.583563 | .4905876 | 3.23  | 0.001 | .6220287 2.545097  |
| 12       | 1.623576 | .5372947 | 3.02  | 0.003 | .5704977 2.676654  |
| program#day |       |           |       |       |                      |
| 2 4      | -.169034 | .4423472 | -0.38 | 0.702 | -1.036019 .6979505  |
| 2 6      | .2113012 | .4982385 | 0.42  | 0.671 | -.7652283 1.187831 |
| 2 8      | -.1299763 | .6524813 | -0.20 | 0.842 | -1.408816 1.148864 |
| 2 12     | .3212829 | .7306782 | 0.44  | 0.660 | -.11082 1.753386   |
| _cons    | 79.6875 | .7560448 | 105.40| 0.000 | 78.20568 81.16932  |

Random-effects Parameters

<table>
<thead>
<tr>
<th>id:</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>(empty)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual: Unstructured</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var(e0)</td>
<td>9.14566</td>
<td>2.126248</td>
<td>5.79858 14.42475</td>
</tr>
<tr>
<td>var(e4)</td>
<td>11.87114</td>
<td>2.761219</td>
<td>7.524948 18.72757</td>
</tr>
<tr>
<td>var(e6)</td>
<td>10.06571</td>
<td>2.348863</td>
<td>6.371091 15.90284</td>
</tr>
<tr>
<td>var(e8)</td>
<td>13.22464</td>
<td>3.113921</td>
<td>8.335981 20.98026</td>
</tr>
<tr>
<td>var(e12)</td>
<td>13.16909</td>
<td>3.167347</td>
<td>8.219208 21.09995</td>
</tr>
<tr>
<td>cov(e0,e4)</td>
<td>9.625236</td>
<td>2.313971</td>
<td>5.054659 14.19581</td>
</tr>
<tr>
<td>cov(e0,e6)</td>
<td>8.489043</td>
<td>2.106377</td>
<td>4.36062 12.61747</td>
</tr>
<tr>
<td>cov(e0,e8)</td>
<td>9.280414</td>
<td>2.369554</td>
<td>4.636173 13.92465</td>
</tr>
<tr>
<td>cov(e0,e12)</td>
<td>8.898006</td>
<td>2.348243</td>
<td>4.295535 13.50048</td>
</tr>
<tr>
<td>cov(e4,e6)</td>
<td>10.49185</td>
<td>2.492529</td>
<td>5.606578 15.37711</td>
</tr>
<tr>
<td>cov(e4,e8)</td>
<td>11.89787</td>
<td>2.848751</td>
<td>6.314421 17.48132</td>
</tr>
<tr>
<td>cov(e4,e12)</td>
<td>11.28344</td>
<td>2.805027</td>
<td>5.785689 16.78119</td>
</tr>
<tr>
<td>cov(e6,e8)</td>
<td>11.0507</td>
<td>2.649988</td>
<td>5.862697 16.23677</td>
</tr>
<tr>
<td>cov(e6,e12)</td>
<td>10.5006</td>
<td>2.590278</td>
<td>5.423748 15.57745</td>
</tr>
<tr>
<td>cov(e8,e12)</td>
<td>12.4091</td>
<td>3.010796</td>
<td>6.508051 18.31016</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: chi2(14) = 314.67 Prob > chi2 = 0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on
the boundary of the parameter space. If this is not true, then the
reported test is conservative.

Because we are using the variable id only to group the repeated measurements and not to introduce
random effects at the subject level, we use the noconstant option to omit any subject-level effects.
The unstructured covariance matrix is the most general and contains many parameters. In this example, we estimate a distinct residual variance for each day and a distinct covariance for each pair of days.

That there is positive covariance between all pairs of measurements is evident, but what is not as evident is whether the covariances may be more parsimoniously represented. One option would be to explore whether the correlation diminishes as the time gap between strength measurements increases and whether it diminishes systematically. Given the irregularity of the time intervals, an exponential structure would be more appropriate than, say, an AR or MA structure.

```
. estimates store unstructured
. mixed strength i.program##i.day || id:, noconstant
> residuals(exponential, t(day)) nolog nofetable
```

Mixed-effects ML regression

<table>
<thead>
<tr>
<th>Number of obs</th>
<th>173</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of groups</td>
<td>37</td>
</tr>
<tr>
<td>Obs per group: min</td>
<td>3</td>
</tr>
<tr>
<td>avg</td>
<td>4.7</td>
</tr>
<tr>
<td>max</td>
<td>5</td>
</tr>
</tbody>
</table>

Wald chi2(9) = 36.77

Log likelihood = -307.83324

Prob > chi2 = 0.0000

<table>
<thead>
<tr>
<th>Random-effects Parameters</th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>id: (empty)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual: Exponential</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rho</td>
<td>.9786462</td>
<td>.0051238</td>
<td>.9659207 .9866854</td>
</tr>
<tr>
<td>var(e)</td>
<td>11.22349</td>
<td>2.338371</td>
<td>7.460765 16.88389</td>
</tr>
</tbody>
</table>

LR test vs. linear regression: chi2(1) = 292.17

Prob > chi2 = 0.0000

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the parameter space. If this is not true, then the reported test is conservative.

In the above example, we suppressed displaying the main regression parameters because they did not differ much from those of the previous model. While the unstructured model estimated 15 variance–covariance parameters, the exponential model claims to get the job done with just 2, a fact that is not disputed by an LR test comparing the two nested models (at least not at the 0.01 level).

```
. lrtest unstructured
```

Likelihood-ratio test

<table>
<thead>
<tr>
<th>LR chi2(13) = 22.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Assumption: . nested in unstructured)</td>
</tr>
</tbody>
</table>

Prob > chi2 = 0.0481

Note: The reported degrees of freedom assumes the null hypothesis is not on the boundary of the parameter space. If this is not true, then the reported test is conservative.
Crossed-effects models

Not all mixed models contain nested levels of random effects.

Example 10

Returning to our longitudinal analysis of pig weights, suppose that instead of (5) we wish to fit

$$\text{weight}_{ij} = \beta_0 + \beta_1 \text{week}_{ij} + u_i + v_j + \epsilon_{ij}$$  \hspace{1cm} (8)

for the $i = 1, \ldots, 9$ weeks and $j = 1, \ldots, 48$ pigs and

$$u_i \sim N(0, \sigma_u^2); \quad v_j \sim N(0, \sigma_v^2); \quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

all independently. Both (5) and (8) assume an overall population-average growth curve $\beta_0 + \beta_1 \text{week}$ and a random pig-specific shift.

The models differ in how week enters into the random part of the model. In (5), we assume that the effect due to week is linear and pig specific (a random slope); in (8), we assume that the effect due to week, $u_i$, is systematic to that week and common to all pigs. The rationale behind (8) could be that, assuming that the pigs were measured contemporaneously, we might be concerned that week-specific random factors such as weather and feeding patterns had significant systematic effects on all pigs.

Model (8) is an example of a two-way crossed-effects model, with the pig effects $v_j$ being crossed with the week effects $u_i$. One way to fit such models is to consider all the data as one big cluster, and treat the $u_i$ and $v_j$ as a series of $9 + 48 = 57$ random coefficients on indicator variables for week and pig. In the notation of (2),

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_9 \\ v_1 \\ \vdots \\ v_{48} \end{bmatrix} \sim N(0, \mathbf{G}); \quad \mathbf{G} = \begin{bmatrix} \sigma_u^2 \mathbf{I}_9 & 0 \\ 0 & \sigma_v^2 \mathbf{I}_{48} \end{bmatrix}$$

Because $\mathbf{G}$ is block diagonal, it can be represented in \texttt{mixed} as repeated-level equations. All we need is an identification variable to identify all the observations as one big group and a way to tell \texttt{mixed} to treat week and pig as factor variables (or equivalently, as two sets of overparameterized indicator variables identifying weeks and pigs, respectively). \texttt{mixed} supports the special group designation _all for the former and the R \texttt{varname} notation for the latter.
. use http://www.stata-press.com/data/r13/pig, clear
(Longitudinal analysis of pig weights)
. mixed weight week || _all: R.week || _all: R.id
Performing EM optimization:
Performing gradient-based optimization:
Iteration 0:  log likelihood = -1013.824
Iteration 1:  log likelihood = -1013.824
Computing standard errors:
Mixed-effects ML regression  Number of obs   =  432
Group variable: _all  Number of groups =   1
Obs per group: min   =  432
         avg   =  432.0
         max   =  432
Wald chi2(1)  =  13258.28
Log likelihood = -1013.824  Prob > chi2    =  0.0000

| Coef.   | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---------|-----------|------|------|----------------------|
| weight  | 6.209896  | .0539313 | 115.14 | 0.000 | 6.104192 6.315599 |
| _cons   | 19.35561  | .6333982 | 30.56  | 0.000 | 18.11418 20.59705 |

Random-effects Parameters

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>var(R.week)</td>
<td>.0849874</td>
<td>.0868856</td>
</tr>
<tr>
<td>var(R.id)</td>
<td>14.83623</td>
<td>3.126142</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>4.297328</td>
<td>.3134404</td>
</tr>
</tbody>
</table>

LR test vs. linear regression:  chi2(2)  =  474.85  Prob > chi2    =  0.0000
Note: LR test is conservative and provided only for reference.

 Thus we estimate $\hat{\sigma}^2_u = 0.08$ and $\hat{\sigma}^2_v = 14.84$. Both (5) and (8) estimate a total of five parameters: two fixed effects and three variance components. The models, however, are not nested within each other, which precludes the use of an LR test to compare both models. Refitting model (5) and looking at the Akaike information criteria values by using estimates stats,

 . quietly mixed weight week || id:week
 . estimates stats crossed.

 Akaike’s information criterion and Bayesian information criterion

<table>
<thead>
<tr>
<th>Model</th>
<th>Obs</th>
<th>ll(null)</th>
<th>ll(model)</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>crossed</td>
<td>432</td>
<td>-1013.824</td>
<td>2037.648</td>
<td>5</td>
<td>2057.99</td>
<td></td>
</tr>
<tr>
<td></td>
<td>432</td>
<td>-869.0383</td>
<td>1748.077</td>
<td>5</td>
<td>1768.419</td>
<td></td>
</tr>
</tbody>
</table>

Note: N=Obs used in calculating BIC; see [R] BIC note
definitely favors model (5). This finding is not surprising given that our rationale behind (8) was somewhat fictitious. In our estimates stats output, the values of ll(null) are missing. mixed does not fit a constant-only model as part of its usual estimation of the full model, but you can use mixed to fit a constant-only model directly, if you wish.
The \texttt{R.varname} notation is equivalent to giving a list of overparameterized (none dropped) indicator variables for use in a random-effects specification. When you specify \texttt{R.varname}, \texttt{mixed} handles the calculations internally rather than creating the indicators in the data. Because the set of indicators is overparameterized, \texttt{R.varname} implies \texttt{noconstant}. You can include factor variables in the fixed-effects specification by using standard methods; see \cite{U} \textbf{11.4.3 Factor variables}. However, random-effects equations support only the \texttt{R.varname} factor specification. For more complex factor specifications (such as interactions) in random-effects equations, use \texttt{generate} to form the variables manually, as we demonstrated in \textbf{example 6}.

\section*{Technical note}

Although we were able to fit the crossed-effects model (8), it came at the expense of increasing the column dimension of our random-effects design from 2 in model (5) to 57 in model (8). Computation time and memory requirements grow (roughly) quadratically with the dimension of the random effects. As a result, fitting such crossed-effects models is feasible only when the total column dimension is small to moderate.

Reexamining model (8), we note that if we drop $u_i$, we end up with a model equivalent to (4), meaning that we could have fit (4) by typing

\begin{verbatim}
   . mixed weight week || _all: R.id
\end{verbatim}

instead of

\begin{verbatim}
   . mixed weight week || id:
\end{verbatim}

as we did when we originally fit the model. The results of both estimations are identical, but the latter specification, organized at the cluster (pig) level with random-effects dimension 1 (a random intercept) is much more computationally efficient. Whereas with the first form we are limited in how many pigs we can analyze, there is no such limitation with the second form.

Furthermore, we fit model (8) by using

\begin{verbatim}
   . mixed weight week || _all: R.week || _all: R.id
\end{verbatim}

as a direct way to demonstrate the \texttt{R.} notation. However, we can technically treat pigs as nested within the \texttt{_all} group, yielding the equivalent and more efficient (total column dimension 10) way to fit (8):

\begin{verbatim}
   . mixed weight week || _all: R.week || id:
\end{verbatim}

We leave it to you to verify that both produce identical results. See Rabe-Hesketh and Skrondal (2012) for additional techniques to make calculations more efficient in more complex models.

\section*{Example 11}

As another example of how the same model may be fit in different ways by using \texttt{mixed} (and as a way to demonstrate \texttt{covariance(exchangeable)}), consider the three-level model used in \textbf{example 4}:

\begin{equation*}
   y_{jk} = X_{jk}\beta + u_k^{(3)} + u_{jk}^{(2)} + \epsilon_{jk}
\end{equation*}
where $y_{jk}$ represents the logarithms of gross state products for the $n_{jk} = 17$ observations from state $j$ in region $k$, $X_{jk}$ is a set of regressors, $u_{k}^{(3)}$ is a random intercept at the region level, and $u_{jk}^{(2)}$ is a random intercept at the state (nested within region) level. We assume that $u_{k}^{(3)} \sim N(0, \sigma_3^2)$ and $u_{jk}^{(2)} \sim N(0, \sigma_2^2)$ independently. Define

$$v_k = \begin{bmatrix} u_{k}^{(3)} + u_{1k}^{(2)} \\ u_{k}^{(3)} + u_{2k}^{(2)} \\ \vdots \\ u_{k}^{(3)} + u_{M_k,k}^{(2)} \end{bmatrix}$$

where $M_k$ is the number of states in region $k$. Making this substitution, we can stack the observations for all the states within region $k$ to get

$$y_k = X_k \beta + Z_k v_k + \epsilon_k$$

where $Z_k$ is a set of indicators identifying the states within each region; that is,

$$Z_k = I_{M_k} \otimes J_{17}$$

for a $k$-column vector of 1s $J_k$, and

$$\Sigma = \text{Var}(v_k) = \begin{bmatrix} \sigma_3^2 + \sigma_2^2 & \sigma_3^2 & \cdots & \sigma_3^2 \\ \sigma_3^2 & \sigma_3^2 + \sigma_2^2 & \cdots & \sigma_3^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_3^2 & \sigma_3^2 & \sigma_3^2 & \sigma_3^2 + \sigma_2^2 \end{bmatrix}_{M_k \times M_k}$$

Because $\Sigma$ is an exchangeable matrix, we can fit this alternative form of the model by specifying the exchangeable covariance structure.
. use http://www.stata-press.com/data/r13/productivity
(Public Capital Productivity)
. mixed gsp private emp hwy water other unemp || region: R.state,
> cov(exchangeable)
(output omitted)
Mixed-effects ML regression
Number of obs = 816
Group variable: region
Number of groups = 9
Obs per group: min = 51
avg = 90.7
max = 136
Log likelihood = 1430.5017
Wald chi2(6) = 18829.06
Prob > chi2 = 0.0000

gsp Coef. Std. Err. z P>|z| [95% Conf. Interval]
private .2671484 .0212591 12.57 0.000 .2254813 .3088154
emp .7540721 .0261868 28.80 0.000 .7027468 .8053973
hwy .0709767 .023041 3.08 0.002 .0258172 .1161363
water .0761187 .0139248 5.47 0.000 .0488266 .1034109
other -.0999955 .0169366 -5.90 0.000 -.1331907 -.0668004
unemp -.0058983 .0009031 -6.53 0.000 -.0076684 -.0041282
_cons 2.128823 .1543855 13.79 0.000 1.826233 2.431413

Random-effects Parameters
Estimate Std. Err. [95% Conf. Interval]
region: Exchangeable
var(R.state) .0077263 .0017926 .0049032 .0121749
cov(R.state) .0014506 .0012995 -.0010963 .0039975

var(Residual) .0013461 .0000689 .0012176 .0014882

LR test vs. linear regression: chi2(2) = 1154.73 Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.

The estimates of the fixed effects and their standard errors are equivalent to those from example 4, and remapping the variance components from \((\sigma_3^2 + \sigma_2^2, \sigma_3^2, \sigma_2^2)\), as displayed here, to \((\sigma_3^2, \sigma_2^2, \sigma_2^2)\), as displayed in example 4, will show that they are equivalent as well.

Of course, given the discussion in the previous technical note, it is more efficient to fit this model as we did originally, as a three-level model.

---

Diagnosing convergence problems

Given the flexibility of mixed-effects models, you will find that some models fail to converge when used with your data; see Diagnosing convergence problems in [ME] me for advice applicable to mixed-effects models in general.

In unweighted LME models with independent and homoskedastic residuals, one useful way to diagnose problems of nonconvergence is to rely on the EM algorithm (Dempster, Laird, and Rubin 1977), normally used by mixed only as a means of refining starting values. The advantages of EM are that it does not require a Hessian calculation, each successive EM iteration will result in a larger likelihood, iterations can be calculated quickly, and iterations will quickly bring parameter estimates into a neighborhood of the solution. The disadvantages of EM are that, once in a neighborhood of the
solution, it can be slow to converge, if at all, and EM provides no facility for estimating standard errors of the estimated variance components. One useful property of EM is that it is always willing to provide a solution if you allow it to iterate enough times, if you are satisfied with being in a neighborhood of the optimum rather than right on the optimum, and if standard errors of variance components are not crucial to your analysis.

If you encounter a nonconvergent model, try using the `emonly` option to bypass gradient-based optimization. Use `emiterate(#)` to specify the maximum number of EM iterations, which you will usually want to set much higher than the default of 20. If your EM solution shows an estimated variance component that is near 0, a ridge is formed by an interval of values near 0, which produces the same likelihood and looks equally good to the optimizer. In this case, the solution is to drop the offending variance component from the model.

Survey data

Multilevel modeling of survey data is a little different from standard modeling in that weighted sampling can take place at multiple levels in the model, resulting in multiple sampling weights. Most survey datasets, regardless of the design, contain one overall inclusion weight for each observation in the data. This weight reflects the inverse of the probability of ultimate selection, and by “ultimate” we mean that it factors in all levels of clustered sampling, corrections for noninclusion and oversampling, poststratification, etc.

For simplicity, in what follows assume a simple two-stage sampling design where groups are randomly sampled and then individuals within groups are sampled. Also assume that no additional weight corrections are performed; that is, sampling weights are simply the inverse of the probability of selection. The sampling weight for observation $i$ in cluster $j$ in our two-level sample is then $w_{ij} = 1/\pi_{ij}$, where $\pi_{ij}$ is the probability that observation $i,j$ is selected. If you were performing a standard analysis such as OLS regression with `regress`, you would simply use a variable holding $w_{ij}$ as your `pweight` variable, and the fact that it came from two levels of sampling would not concern you. Perhaps you would type `vce(cluster groupvar)` where `groupvar` identifies the top-level groups to get standard errors that control for correlation within these groups, but you would still use only a single weight variable.

Now take these same data and fit a two-level model with `mixed`. As seen in (14) in Methods and formulas later in this entry, it is not sufficient to use the single sampling weight $w_{ij}$, because weights enter into the log likelihood at both the group level and the individual level. Instead, what is required for a two-level model under this sampling design is $w_j$, the inverse of the probability that group $j$ is selected in the first stage, and $w_{i|j}$, the inverse of the probability that individual $i$ from group $j$ is selected at the second stage conditional on group $j$ already being selected. It simply will not do to just use $w_{ij}$ without making any assumptions about $w_j$.

Given the rules of conditional probability, $w_{ij} = w_j w_{i|j}$. If your dataset has only $w_{ij}$, then you will need to either assume equal probability sampling at the first stage ($w_j = 1$ for all $j$) or find some way to recover $w_j$ from other variables in your data; see Rabe-Hesketh and Skrondal (2006) and the references therein for some suggestions on how to do this, but realize that there is little yet known about how well these approximations perform in practice.

What you really need to fit your two-level model are data that contain $w_j$ in addition to either $w_{ij}$ or $w_{i|j}$. If you have $w_{ij}$—that is, the unconditional inclusion weight for observation $i,j$—then you need to either divide $w_{ij}$ by $w_j$ to obtain $w_{i|j}$ or rescale $w_{ij}$ so that its dependence on $w_j$ disappears. If you already have $w_{i|j}$, then rescaling becomes optional (but still an important decision to make).

Weight rescaling is not an exact science, because the scale of the level-one weights is at issue regardless of whether they represent $w_{ij}$ or $w_{i|j}$: because $w_{ij}$ is unique to group $j$, the group-to-group
magnitudes of these weights need to be normalized so that they are “consistent” from group to group. This is in stark contrast to a standard analysis, where the scale of sampling weights does not factor into estimation, instead only affecting the estimate of the total population size.

mixed offers three methods for standardizing weights in a two-level model, and you can specify which method you want via the pwscale() option. If you specify pwscale(size), then the \( w_{ij} \) (or \( w_{ij} \), it does not matter) are scaled to sum to the cluster size \( n_j \). Method pwscale(effective) adds in a dependence on the sum of the squared weights so that level-one weights sum to the “effective” sample size. Just like pwscale(size), pwscale(effective) also behaves the same whether you have \( w_{ij} \) or \( w_{ij} \), and so it can be used with either.

Although both pwscale(size) and pwscale(effective) leave \( w_j \) untouched, the pwscale(gk) method is a little different in that 1) it changes the weights at both levels and 2) it does assume you have \( w_{ij} \) for level-one weights and not \( w_{ij} \) (if you have the latter, then first divide by \( w_j \)). Using the method of Graubard and Korn (1996), it sets the weights at the group level (level two) to the cluster averages of the products of both level weights (this product being \( w_{ij} \)). It then sets the individual weights to 1 everywhere; see Methods and formulas for the computational details of all three methods.

Determining which method is “best” is a tough call and depends on cluster size (the smaller the clusters, the greater the sensitivity to scale), whether the sampling is informative (that is, the sampling weights are correlated with the residuals), whether you are interested primarily in regression coefficients or in variance components, whether you have a simple random-intercept model or a more complex random-coefficients model, and other factors; see Rabe-Hesketh and Skrondal (2006), Carle (2009), and Pfeffermann et al. (1998) for some detailed advice. At the very least, you want to compare estimates across all three scaling methods (four, if you add no scaling) and perform a sensitivity analysis.

If you choose to rescale level-one weights, it does not matter whether you have \( w_{ij} \) or \( w_{ij} \). For the pwscale(size) and pwscale(effective) methods, you get identical results, and even though pwscale(gk) assumes \( w_{ij} \), you can obtain this as \( w_{ij} = w_{ij}/w_j \) before proceeding.

If you do not specify pwscale(), then no scaling takes place, and thus at a minimum, you need to make sure you have \( w_{ij} \) in your data and not \( w_{ij} \).

Example 12

Rabe-Hesketh and Skrondal (2006) analyzed data from the 2000 Programme for International Student Assessment (PISA) study on reading proficiency among 15-year-old American students, as performed by the Organisation for Economic Co-operation and Development (OECD). The original study was a three-stage cluster sample, where geographic areas were sampled at the first stage, schools at the second, and students at the third. Our version of the data does not contain the geographic-areas variable, so we treat this as a two-stage sample where schools are sampled at the first stage and students at the second.
For student \( i \) in school \( j \), where the variable `id_school` identifies the schools, the variable `w_fstuwt` is a student-level overall inclusion weight \( w_{ij} \), adjusted for noninclusion and nonparticipation of students, and the variable `wnrschbw` is the school-level weight \( w_j \) adjusted for oversampling of schools with more minority students. The weight adjustments do not interfere with the methods prescribed above, and thus we can treat the weight variables simply as \( w_{ij} \) and \( w_j \), respectively.

Rabe-Hesketh and Skrondal (2006) fit a two-level logistic model for passing a reading proficiency threshold. We fit a two-level linear random-intercept model for socioeconomic index. Because we have \( w_{ij} \) and not \( w_{i|j} \), we rescale using `pweights(size)` and thus obtain results as if we had \( w_{i|j} \).
mixed isei female high_school college one_for both_for test_lang
> [pw=f fstuwt] || id_school:, pweight(wnrschbw) pwscale(size)

(output omitted)

Mixed-effects regression
Group variable: id_school

|                | Coef.  | Std. Err. | z     | P>|z|  | [95% Conf. Interval] |
|----------------|--------|-----------|-------|-------|----------------------|
| female         | .59379 | .8732886  | 0.68  | 0.497 | -1.117824 2.305404   |
| high_school    | 6.410618 | 1.500337 | 4.27 | 0.000 | 3.470011 9.351224   |
| college        | 19.39494 | 2.121145 | 9.14 | 0.000 | 15.23757 23.55231   |
| one_for        | -.9584613 | 0.503337 | -1.89 | 0.060 | -2.906925 0.990004   |
| both_for       | -.2021101 | 2.326333 | -0.09 | 0.922 | -4.761633 4.357413   |
| test_lang      | 2.519539 | 2.393165 | 1.05 | 0.292 | -2.170978 7.210056   |
| cons           | 28.10788 | 2.435712 | 11.54 | 0.000 | 23.33397 32.88179   |

Random-effects Parameters

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>id_school: Identity var(_cons)</td>
<td>34.69374</td>
<td>8.574865</td>
<td>21.37318 56.31617</td>
</tr>
<tr>
<td>var(Residual)</td>
<td>218.7382</td>
<td>11.22111</td>
<td>197.8147 241.8748</td>
</tr>
</tbody>
</table>

Notes:
1. We specified the level-one weights using standard Stata weight syntax, that is, [pw=f fstuwt].
2. We specified the level-two weights via the pweight(wnrschbw) option as part of the random-effects specification for the id_school level. As such, it is treated as a school-level weight. Accordingly, wnrschbw needs to be constant within schools, and mixed did check for that before estimating.
3. Because our level-one weights are unconditional, we specified pwscale(size) to rescale them.
4. As is the case with other estimation commands in Stata, standard errors in the presence of sampling weights are robust.
5. Robust standard errors are clustered at the top level of the model, and this will always be true unless you specify vce(cluster clustvar), where clustvar identifies an even higher level of grouping.

As a form of sensitivity analysis, we compare the above with scaling via pwscale(gk). Because pwscale(gk) assumes $w_{ij}$, you want to first divide $w_{ij}$ by $w_j$. But you can handle that within the weight specification itself.
The results are somewhat similar to before, which is good news from a sensitivity standpoint. Note that we specified \([pw=w_{fstwtw}/wnrschbw]\) and thus did the conversion from \(w_{ij}\) to \(w_{i|j}\) within our call to `mixed`.

We close this section with a bit of bad news. Although weight rescaling and the issues that arise have been well studied for two-level models, as pointed out by Carle (2009), “... a best practice for scaling weights across multiple levels has yet to be advanced.” As such, `pwscale()` is currently supported only for two-level models. If you are fitting a higher-level model with survey data, you need to make sure your sampling weights are conditional on selection at the previous stage and not overall inclusion weights, because there is currently no rescaling option to fall back on if you do not.
mixed stores the following in e():

Scalars

- `e(N)` number of observations
- `e(k)` number of parameters
- `e(k_f)` number of fixed-effects parameters
- `e(k_r)` number of random-effects parameters
- `e(k_rs)` number of variances
- `e(k_rc)` number of covariances
- `e(k_res)` number of residual-error parameters
- `e(N_clust)` number of clusters
- `e(nrgroups)` number of residual-error by() groups
- `e(ar_p)` AR order of residual errors, if specified
- `e(ma_q)` MA order of residual errors, if specified
- `e(res_order)` order of residual-error structure, if appropriate
- `e(df_m)` model degrees of freedom
- `e(ll)` log (restricted) likelihood
- `e(chi2)` \( \chi^2 \)
- `e(p)` significance
- `e(l1_c)` log likelihood, comparison model
- `e(chi2_c)` \( \chi^2 \), comparison model
- `e(df_c)` degrees of freedom, comparison model
- `e(p_c)` significance, comparison model
- `e(rank)` rank of \( e(V) \)
- `e(rc)` return code
- `e(converged)` 1 if converged, 0 otherwise

Macros

- `e(cmd)` `mixed`
- `e(cmdline)` command as typed
- `e(depvar)` name of dependent variable
- `e(wtype)` weight type (first-level weights)
- `e(wexp)` weight expression (first-level weights)
- `e(fweight)` weight expression for \( k \)th highest level, if specified
- `e(pweight)` weight expression for \( k \)th highest level, if specified
- `e(ivars)` grouping variables
- `e(title)` title in estimation output
- `e(redim)` random-effects dimensions
- `e(vartypes)` variance-structure types
- `e(revars)` random-effects covariates
- `e(resopt)` residuals() specification, as typed
- `e(rstructure)` residual-error structure
- `e(rstructlab)` residual-error structure output label
- `e(rbyvar)` residual-error by() variable, if specified
- `e(rglabels)` residual-error by() groups labels
- `e(psweight)` sampling-weight scaling method
- `e(timevar)` residual-error t() variable, if specified
- `e(chi2type)` Wald; type of model \( \chi^2 \) test
- `e(clustvar)` name of cluster variable
- `e(vcse)` `vcetype` specified in `vce()`
- `e(vcetype)` title used to label Std. Err.
- `e(method)` ML or REML
- `e(opt)` type of optimization
- `e(optmetric)` matsqrt or matlog; random-effects matrix parameterization
- `e(emonly)` emonly, if specified
- `e(ml_method)` type of ml method
- `e(technique)` maximization technique
- `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
Methods and formulas

As given by (1), in the absence of weights we have the linear mixed model

\[ y = X\beta + Zu + \epsilon \]

where \( y \) is the \( n \times 1 \) vector of responses, \( X \) is an \( n \times p \) design/covariate matrix for the fixed effects \( \beta \), and \( Z \) is the \( n \times q \) design/covariate matrix for the random effects \( u \). The \( n \times 1 \) vector of errors \( \epsilon \) is for now assumed to be multivariate normal with mean 0 and variance matrix \( \sigma^2 \epsilon I_n \). We also assume that \( u \) has variance–covariance matrix \( G \) and that \( u \) is orthogonal to \( \epsilon \) so that

\[
\text{Var} \begin{bmatrix} u \\ \epsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & \sigma^2 \epsilon I_n \end{bmatrix}
\]

Considering the combined error term \( Zu + \epsilon \), we see that \( y \) is multivariate normal with mean \( X\beta \) and \( n \times n \) variance–covariance matrix

\[ V = ZGZ' + \sigma^2 \epsilon I_n \]

Defining \( \theta \) as the vector of unique elements of \( G \) results in the log likelihood

\[ L(\beta, \theta, \sigma^2 \epsilon) = -\frac{1}{2} \left\{ n \log(2\pi) + \log |V| + (y - X\beta)'V^{-1}(y - X\beta) \right\} \] (9)

which is maximized as a function of \( \beta \), \( \theta \), and \( \sigma^2 \epsilon \). As explained in chapter 6 of Searle, Casella, and McCulloch (1992), considering instead the likelihood of a set of linear contrasts \( Ky \) that do not depend on \( \beta \) results in the restricted log likelihood

\[ L_R(\beta, \theta, \sigma^2 \epsilon) = L(\beta, \theta, \sigma^2 \epsilon) - \frac{1}{2} \log |X'V^{-1}X| \] (10)

Given the high dimension of \( V \), however, the log-likelihood and restricted log-likelihood criteria are not usually computed by brute-force application of the above expressions. Instead, you can simplify the problem by subdividing the data into independent clusters (and subclusters if possible) and using matrix decomposition methods on the smaller matrices that result from treating each cluster one at a time.

Consider the two-level model described previously in (2),

\[ y_j = X_j\beta + Z_ju_j + \epsilon_j \]

for \( j = 1, \ldots, M \) clusters with cluster \( j \) containing \( n_j \) observations, with \( \text{Var}(u_j) = \Sigma \), a \( q \times q \) matrix.
Efficient methods for computing (9) and (10) are given in chapter 2 of Pinheiro and Bates (2000). Namely, for the two-level model, define \( \Delta \) to be the Cholesky factor of \( \sigma_e^2 \Sigma^{-1} \), such that \( \sigma_e^2 \Sigma^{-1} = \Delta' \Delta \). For \( j = 1, \ldots, M \), decompose

\[
\begin{bmatrix}
Z_j \\
\Delta
\end{bmatrix} = Q_j \begin{bmatrix}
R_{11j} \\
0
\end{bmatrix}
\]

by using an orthogonal-triangular (QR) decomposition, with \( Q_j \) a \((n_j + q)\)-square matrix and \( R_{11j} \) a \(q\)-square matrix. We then apply \( Q_j \) as follows:

\[
\begin{bmatrix}
R_{10j} \\
R_{00j}
\end{bmatrix} = Q_j' \begin{bmatrix}
X_j \\
0
\end{bmatrix} ; \quad \begin{bmatrix}
c_{1j} \\
c_{0j}
\end{bmatrix} = Q_j' \begin{bmatrix}
y_j \\
0
\end{bmatrix}
\]

Stack the \( R_{00j} \) and \( c_{0j} \) matrices, and perform the additional QR decomposition

\[
\begin{bmatrix}
R_{001} & c_{01} \\
: & : \\
R_{00M} & c_{0M}
\end{bmatrix} = Q_0 \begin{bmatrix}
R_{00} & c_0 \\
0 & c_1
\end{bmatrix}
\]

Pinheiro and Bates (2000) show that ML estimates of \( \beta, \sigma_e^2 \), and \( \Delta \) (the unique elements of \( \Delta \), that is) are obtained by maximizing the profile log likelihood (profiled in \( \Delta \))

\[
L(\Delta) = \frac{n}{2} \left\{ \log n - \log(2\pi) - 1 \right\} - n \log ||c_1|| + \sum_{j=1}^{M} \log \left| \frac{\det(\Delta)}{\det(R_{11j})} \right| \quad (11)
\]

where \( || \cdot || \) denotes the 2-norm. Following this maximization with

\[
\hat{\beta} = R_{00}^{-1} c_0; \quad \hat{\sigma}_e^2 = n^{-1} ||c_1||^2 \quad (12)
\]

REML estimates are obtained by maximizing

\[
L_R(\Delta) = \frac{n - p}{2} \left\{ \log(n - p) - \log(2\pi) - 1 \right\} - (n - p) \log ||c_1||
\]

\[
- \log |\det(R_{00})| + \sum_{j=1}^{M} \log \left| \frac{\det(\Delta)}{\det(R_{11j})} \right| \quad (13)
\]

followed by

\[
\hat{\beta} = R_{00}^{-1} c_0; \quad \hat{\sigma}_e^2 = (n - p)^{-1} ||c_1||^2
\]

For numerical stability, maximization of (11) and (13) is not performed with respect to the unique elements of \( \Delta \) but instead with respect to the unique elements of the matrix square root (or matrix logarithm if the matlog option is specified) of \( \Sigma/\sigma_e^2 \); define \( \gamma \) to be the vector containing these elements.

Once maximization with respect to \( \gamma \) is completed, \( (\gamma, \sigma_e^2) \) is reparameterized to \( \{\alpha, \log(\sigma_e^2)\} \), where \( \alpha \) is a vector containing the unique elements of \( \Sigma \), expressed as logarithms of standard deviations for the diagonal elements and hyperbolic arctangents of the correlations for off-diagonal elements. This last step is necessary 1) to obtain a joint variance–covariance estimate of the elements of \( \Sigma \) and \( \sigma_e^2 \); 2) to obtain a parameterization under which parameter estimates can be interpreted individually, rather than as elements of a matrix square root (or logarithm); and 3) to parameterize these elements such that their ranges each encompass the entire real line.
Obtaining a joint variance–covariance matrix for the estimated \{\alpha, \log(\sigma_\epsilon)\} requires the evaluation of the log likelihood (or log-restricted likelihood) with only \beta profiled out. For ML, we have

\[
L^* \{\alpha, \log(\sigma_\epsilon)\} = L \{\Delta(\alpha, \sigma_\epsilon^2), \sigma_\epsilon^2\} = -\frac{n}{2} \log(2\pi \sigma_\epsilon^2) - \frac{||c_1||^2}{2\sigma_\epsilon^2} + \sum_{j=1}^{M} \frac{\det(\Delta)}{\det(R_{11j})}
\]

with the analogous expression for REML.

The variance–covariance matrix of \hat{\beta} is estimated as

\[
\hat{\text{Var}}(\hat{\beta}) = \hat{\sigma}_\epsilon^2 R_{00}^{-1} (R_{00}^{-1})'
\]

but this does not mean that \hat{\text{Var}}(\hat{\beta}) is identical under both ML and REML because \(R_{00}\) depends on \(\Delta\). Because \(\hat{\beta}\) is asymptotically uncorrelated with \(\{\hat{\alpha}, \log(\hat{\sigma}_\epsilon)\}\), the covariance of \(\hat{\beta}\) with the other estimated parameters is treated as 0.

Parameter estimates are stored in \(e(b)\) as \(\{\hat{\beta}, \hat{\alpha}, \log(\hat{\sigma}_\epsilon)\}\), with the corresponding (block-diagonal) variance–covariance matrix stored in \(e(V)\). Parameter estimates can be displayed in this metric by specifying the \textit{estmetric} option. However, in \textit{mixed} output, variance components are most often displayed either as variances and covariances or as standard deviations and correlations.

EM iterations are derived by considering the \(u_j\) in (2) as missing data. Here we describe the procedure for maximizing the log likelihood via EM; the procedure for maximizing the restricted log likelihood is similar. The log likelihood for the full data \((y, u)\) is

\[
L_F(\beta, \Sigma, \sigma_\epsilon^2) = \sum_{j=1}^{M} \left\{ \log f_1(y_j|u_j, \beta, \sigma_\epsilon^2) + \log f_2(u_j|\Sigma) \right\}
\]

where \(f_1(\cdot)\) is the density function for multivariate normal with mean \(X_j \beta + Z_j u_j\) and variance \(\sigma_\epsilon^2 I_{n_j}\), and \(f_2(\cdot)\) is the density for multivariate normal with mean 0 and \(q \times q\) covariance matrix \(\Sigma\). As before, we can profile \(\beta\) and \(\sigma_\epsilon^2\) out of the optimization, yielding the following EM iterative procedure:

1. For the current iterated value of \(\Sigma^{(t)}\), fix \(\hat{\beta} = \hat{\beta}(\Sigma^{(t)})\) and \(\hat{\sigma}_\epsilon^2 = \hat{\sigma}_\epsilon^2(\Sigma^{(t)})\) according to (12).
2. Expectation step: Calculate

\[
D(\Sigma) \equiv E \left\{ L_F(\hat{\beta}, \Sigma, \hat{\sigma}_\epsilon^2)|y \right\} = C - \frac{M}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{j=1}^{M} E \left( u_j' \Sigma^{-1} u_j | y \right)
\]

where \(C\) is a constant that does not depend on \(\Sigma\), and the expected value of the quadratic form \(u_j' \Sigma^{-1} u_j\) is taken with respect to the conditional density \(f(\Sigma^{(t)} | y, \hat{\beta}, \hat{\sigma}_\epsilon^2)\).

3. Maximization step: Maximize \(D(\Sigma)\) to produce \(\Sigma^{(t+1)}\).
For general, symmetric $\Sigma$, the maximizer of $D(\Sigma)$ can be derived explicitly, making EM iterations quite fast.

For general, residual-error structures, 

$$\text{Var}(\epsilon_j) = \sigma^2_\epsilon \Lambda_j$$

where the subscript $j$ merely represents that $\epsilon_j$ and $\Lambda_j$ vary in dimension in unbalanced data, the data are first transformed according to

$$y^*_j = \hat{\Lambda}_j^{-1/2} y_j; \quad X^*_j = \hat{\Lambda}_j^{-1/2} X_j; \quad Z^*_j = \hat{\Lambda}_j^{-1/2} Z_j;$$

and the likelihood-evaluation techniques described above are applied to $y^*_j$, $X^*_j$, and $Z^*_j$ instead.

The unique elements of $\Lambda$, $\rho$, are estimated along with the fixed effects and variance components. Because $\sigma^2_\epsilon$ is always estimated and multiplies the entire $\Lambda_j$ matrix, $\hat{\rho}$ is parameterized to take this into account.

In the presence of sampling weights, following Rabe-Hesketh and Skrondal (2006), the weighted log pseudolikelihood for a two-level model is given as

$$L(\beta, \Sigma, \sigma^2_\epsilon) = \sum_{j=1}^{M} w_j \log \left[ \int \exp \left\{ \sum_{i=1}^{n_j} w_{ij} \log f_1(y_{ij} | u_j, \beta, \sigma^2_\epsilon) \right\} f_2(u_j | \Sigma) du_j \right]$$

(14)

where $w_j$ is the inverse of the probability of selection for the $j$th cluster, $w_{ij}$ is the inverse of the conditional probability of selection of individual $i$ given the selection of cluster $j$, and $f_1(\cdot)$ and $f_2(\cdot)$ are the multivariate normal densities previously defined.

Weighted estimation is achieved through incorporating $w_j$ and $w_{ij}$ into the matrix decomposition methods detailed above to reflect replicated clusters for $w_j$ and replicated observations within clusters for $w_{ij}$. Because this estimation is based on replicated clusters and observations, frequency weights are handled similarly.

Rescaling of sampling weights can take one of three available forms:

Under pwscale(size),

$$w^*_{ij} = n_j w_{ij} \left\{ \sum_{i=1}^{n_j} w_{ij} \right\}^{-1}$$

Under pwscale(effective),

$$w^*_{ij} = \left\{ \sum_{i=1}^{n_j} w_{ij} \right\}^{n_j} \left\{ \sum_{i=1}^{n_j} w^2_{ij} \right\}^{-1}$$

Under both the above, $w_j$ remains unchanged. For method pwscale(gk), however, both weights are modified:

$$w^*_j = n_j^{-1} \sum_{i=1}^{n_j} w_{ij} w_j; \quad w^*_{ij} = 1$$

Under ML estimation, robust standard errors are obtained in the usual way (see [P] _robust) with the one distinction being that in multilevel models, robust variances are, at a minimum, clustered at the highest level. This is because given the form of the log likelihood, scores aggregate at the top-level clusters. For a two-level model, scores are obtained as the partial derivatives of $L_j(\beta, \Sigma, \sigma^2_\epsilon)$ with respect to $\{\beta, \alpha, \log(\sigma_\epsilon)\}$, where $L_j$ is the log likelihood for cluster $j$ and $L = \sum_{j=1}^{M} L_j$. Robust variances are not supported under REML estimation because the form of the log restricted likelihood does not lend itself to separation by highest-level clusters.
EM iterations always assume equal weighting and an independent, homoskedastic error structure. As such, with weighted data or when error structures are more complex, EM is used only to obtain starting values.

For extensions to models with three or more levels, see Bates and Pinheiro (1998) and Rabe-Hesketh and Skrondal (2006).

Charles Roy Henderson (1911–1989) was born in Iowa and grew up on the family farm. His education in animal husbandry, animal nutrition, and statistics at Iowa State was interspersed with jobs in the Iowa Extension Service, Ohio University, and the U.S. Army. After completing his PhD, Henderson joined the Animal Science faculty at Cornell. He developed and applied statistical methods in the improvement of farm livestock productivity through genetic selection, with particular focus on dairy cattle. His methods are general and have been used worldwide in livestock breeding and beyond agriculture. Henderson’s work on variance components and best linear unbiased predictions has proved to be one of the main roots of current mixed-model methods.

Acknowledgments

We thank Badi Baltagi of the Department of Economics at Syracuse University and Ray Carroll of the Department of Statistics at Texas A&M University for each providing us with a dataset used in this entry.

References


Also see

[ME] mixed postestimation — Postestimation tools for mixed

[ME] me — Introduction to multilevel mixed-effects models

[Mi] estimation — Estimation commands for use with mi estimate

[SEM] intro 5 — Tour of models (Multilevel mixed-effects models)

[XT] xtrc — Random-coefficients model

[XT] xtreg — Fixed-, between-, and random-effects and population-averaged linear models

[U] 20 Estimation and postestimation commands