

Intro 6 — Panel data and grouped data model features

[Description](#)[Remarks and examples](#)[Also see](#)

Description

In panel data and in other grouped data, observations within the same panel or group are not independent. The `xt` versions of the ERM commands fit models with random effects to account for the within-panel or within-group correlation. `xtregress`, `xteintreg`, `xtprobit`, and `xteoprobit` are explained below.

Remarks and examples

[stata.com](#)

Remarks are presented under the following headings:

Random-effects models that ERMs handle

Random effects can be used with other features of ERMs

Random-effects models that ERMs handle

In [ERM] [Intro 2](#) through [ERM] [Intro 5](#), we discussed models with observation-level errors that are assumed independent and identically distributed.

When we typed

```
. eregress y x1 x2
```

the model fit was

$$y_j = \beta_0 + \beta_1 x1_j + \beta_2 x2_j + e_{j.y}$$

For this model, we could tell a story about how y_j is the college grade point average (GPA) of student j and the error $e_{j.y}$ represents the unobserved factors that influence y_j . Each observation in these data could be a randomly drawn student.

What if we record the semester GPA for each student for eight semesters? In addition to the observation-level error that is represented in $e_{j.y}$, the unobserved characteristics of the student (such as student ability) may have an effect on y_j .

We can adjust our model to include effect of student,

$$y_{ij} = \beta_0 + \beta_1 x1_{ij} + \beta_2 x2_{ij} + u_{i.y} + e_{ij.y}$$

Now the unobserved effect of semester j for student i is $e_{ij.y}$. The unobserved effect of the student is $u_{i.y}$. We treat the effect of student as random.

Some disciplines would refer to this model as a panel-data random-effects model where students are the panels. Others would refer to it as a multilevel model with two levels—the student level and the semester-within-student level—with random intercepts at the student level. Regardless of terminology, the model is the same. We estimate variance parameters for the student-level random effects and for the observation-level errors.

The ERM commands for random-effects models follow the style of Stata's panel-data (`xt`) commands.

If the variable `studentid` identifies the students in our sample and the variable `semester` identifies the semester, we can fit the model above by typing

```
. xtset studentid semester
. xtregress y x1 x2
```

The `xtset` command left information so that `xtregress` recognized `studentid` as the panel identifier.

Although we use the panel-data syntax, these ERM commands do not require a traditional panel dataset where you have repeated time periods within panels.

Let's change our story so that each observation is a different student. We observe GPA only once for each student. However, we still have grouped data because the students are randomly chosen from multiple colleges. Students from the same college may have more in common than students from different colleges. Now our two levels are colleges and students nested within college. College can be specified in `xtset` as the panel identifier. To fit the model, we type

```
. xtset college
. xtregress y x1 x2
```

If we have a different type of outcome, we can use one of the other panel-data ERM commands—`xteintreg`, `xteprobit`, or `xteoprobit`. For instance, if `y` is a binary indicator of whether the student graduated, we would type

```
. xtset college
. xteprobit y x1 x2
```

Ignoring the panel or group structure of your data can lead to inefficient estimates in the linear model and inconsistent estimates for nonlinear models. The `xtregress`, `xteintreg`, `xteprobit`, and `xteoprobit` commands model random effects and provide efficient and consistent estimates.

Random effects can be used with other features of ERMs

We can combine random effects with other features of ERMs, that is, with endogenous covariates, sample selection, and treatment effects.

By typing

```
. xtset college
. xtregress y x1 x2, endogenous(w1 = x1 z1 z2)
```

we would fit a linear regression model for `y` with random effects and an endogenous covariate `w1`. We would have a college-level random effect for `y`. This would be correlated with the college-level random effect for `w1`. The observation-level errors for `y` and `w1` would be correlated as well.

We can similarly include the `select()`, `extreat()`, and `entreat()` options described in [\[ERM\] Intro 4](#) and [\[ERM\] Intro 5](#) to account for sample selection and nonrandom treatment when fitting random-effects models with any of the panel-data ERM commands.

Random effects are included in the main outcome equation as well as in equations for endogenous covariates, sample selection, and endogenous treatment. If we do not believe that the other features should have random effects, we specify the `nore` suboption with `select()`, `extreat()`, or `entreat()`.

Suppose w_1 was high school GPA. The college-level random effect for high school GPA may be negligible. To fit the model with a random effect in the equation for y but not in the equation for the endogenous covariate w_1 , we would type

```
. xtset schoolid
. xtregress y x1 x2, endogenous(w1 = x1 z1 z2, none)
```

Also see

[ERM] [Intro 9](#) — Conceptual introduction via worked example