Models for Longitudinal and Clustered Data

Germán Rodríguez

December 9, 2008, revised December 6, 2012

1 Introduction

The most important assumption we have made in this course is that the observations are *independent*. Situations where this assumption is not appropriate include

- Longitudinal data, where we have repeated observations on each individual, for example on multiple waves of a survey
- Clustered data, where the observations are grouped, for example data on mothers and their children
- Multilevel data, where we have multiple levels of grouping, for example students in classrooms in schools.

This is a large subject worthy of a separate course. In these notes I will review briefly the main approaches to the analysis of this type of data, namely fixed and random-effects models. I will deal with linear models for continuous data in Section 2 and logit models for binary data in section 3. I will describe the models in terms of clustered data, using Y_{ij} to represent the outcome for the j-th member of the i-th group. The same procedures, however, apply to longitudinal data, so Y_{ij} could be the response for the i-th individual on the j-th wave. There is no requirement that all groups have the same number of members or, in the longitudinal case, that all individuals have the same number of measurements.

The Stata section of the course website has relevant logs under 'panel data models', including an analysis of data on verbal IQ and language scores for 2287 children in 131 schools in the Netherlands, and a study of the relationship between low birth weight and participation in the Aid to Families with Dependent Children (AFDC) welfare program using state-level data for 1987 and 1990. For binary data we use an example in the Stata manual. A short do file is included at the end of this document.

2 Continuous Data

Suppose that Y_{ij} is a continuous outcome for the j-th member of group i. We are willing to assume independence across groups, but not within each group. The basic idea is that there may be unobserved group characteristics that affect the outcomes of the individuals in each group. We consider two ways to model these characteristics.

2.1 Fixed Effects

The first model we will consider introduces a separate parameter for each group, so the observations satisfy

$$Y_{ij} = \alpha_i + x'_{ij}\beta + e_{ij} \tag{1}$$

Here α_i is a group-specific parameter representing the effect of unobserved group characteristics, the β are regression coefficients representing the effects of the observed covariates, and the e_{ij} are *independent* error terms, say $e_{ij} \sim N(0, \sigma_e^2)$.

You can think of the α_i as equivalent to introducing a separate dummy variable for each group. It is precisely because we have controlled for (all) group characteristics that we are willing to assume independence of the observations. Unfortunately this implies that we cannot include group-level covariates among the predictors, as they would be collinear with the dummies. Effectively this means that we can control for group characteristics, but we cannot estimate their effects.

This model typically has a large number of parameters, and this causes practical and theoretical problems.

In terms of theory the usual OLS estimator of α_i is consistent as the number of individuals approaches infinity in every group, but is not consistent if the number of groups approaches infinity but the number of individuals per group does not, which is the usual case of interest. Fortunately the OLS estimator of β is consistent in both cases.

On the practical side, introducing a dummy variable for each group may not be feasible when the number of groups is very large. Fortunately it is possible to solve for the OLS estimator of β without having to estimate the α_i's explicitly through a process known as *absorption*.

An alternative is to remove the α_i from the model by differencing or conditioning. This is very easy to do if you have two observations per group, as would be the case for longitudinal data from a two-wave survey. Suppose

Y_{i1} and Y_{i2} follow model (1). The *difference* would then follow the model

$$Y_{i2} - Y_{i1} = (x_{i2} - x_{i1})'\beta + (e_{i2} - e_{i1})$$

which is a linear model with exactly the same regression coefficients as (1). Moreover, because the e_{ij} are independent, so are their differences. This means that we can obtain unbiased estimates of β by simply differencing the Y's and the x's and using ordinary OLS on the differences.

The same idea can be extended to more than two observations per group, and it involves working with a transformation of the data reflecting essentially differences with respect to the group means. The same estimator can also be obtained by working with the conditional distribution of the observations given the group totals $Y_i = \sum_j Y_{ij}$.

Looking at the model in terms of differences shows clearly how it can control for unobserved group characteristics. Suppose the 'true' model includes a group-level predictor z_i with coefficient γ, so

$$Y_{ij} = z_i'\gamma + x_{ij}'\beta + e_{ij}$$

When you difference the y's the term $z_i'\gamma$ drops out. Therefore you can estimate effects of the x's controlling for z even though you haven't observed z! Unfortunately, this also means that in a fixed-effects model we can't estimate γ even if we have observed z_i, as noted earlier.

2.2 Random Effects

An alternative approach writes a model that looks almost identical to the previous one:

$$Y_{ij} = a_i + x_{ij}'\beta + e_{ij} \tag{2}$$

Here a_i is a *random* variable representing a group-specific effect, β is a vector of regression coefficients and the e_{ij} are independent error terms.

You can think of the a_i and e_{ij} as two error terms, one at the level of the group and the other at the level of the individual. As usual with error terms we assign them distributions; specifically we assume that $a_i \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. We also assume that e_{ij} is independent of a_i.

Another way to write the model is by combining the two error terms in one:

$$Y_{ij} = x_{ij}'\beta + u_{ij}$$

where $u_{ij} = a_i + e_{ij}$. This looks like an ordinary regression model, but the errors are not independent. More precisely, they are independent across

groups but not within a subgroup because the u_{ij}'s for members of group i share a_i.

We can write the correlation between any two observations in the same group as

$$\rho = \text{cor}(Y_{ij}, Y_{ij'}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}.$$

a result that follows directly from the usual definition of correlation; the covariance between Y_{ij} and $Y_{ij'}$ is σ_a^2 and the variance of either is $\sigma_a^2 + \sigma_e^2$. This coefficient is often called *the intra-class correlation coefficient.*

Because the variance of the observations has been partitioned into two components these models are also called *variance components models.* The term σ_a^2 represents variation across groups (usually called *between* groups, even if we have more than two) and the term σ_e^2 represents variation *within* groups.

If we were to use OLS estimation in the model of equation (2) we would obtain consistent estimates for the regression coefficients β, but the estimates would not be fully efficient because they do not take into account the covariance structure, and the standard errors would be biased unless they are corrected for clustering.

Fortunately maximum likelihood estimation is pretty straightforward, and yields fully-efficient estimates. We also obtain as by-products estimates of the error variances σ_a^2 and σ_e^2 and the intra-class correlation ρ. (Stata also computes these quantities for fixed-effect models, where they are best viewed as components of the total variance.)

2.3 Fixed Versus Random Effects

There is a lot of confusion regarding fixed and random-effects models. Here are five considerations that may help you decide which approach may be more appropriate for a given problem.

First let us note the obvious, in one case the α_i are fixed but unknown parameters to be estimated (or differenced out of the model), in the other the a_i are random variables and we estimate their distribution, which has mean zero and variance σ_a^2. This distinction leads to one traditional piece of advice: use random effects if you view the groups as a sample from a population, and fixed effects if you are interested in inferences for the specific groups at hand. I find this advice to be the least useful of all. (It is particularly baffling to Bayesians, who view all parameters as random.)

Second, note that the a_i are assumed to be *independent* across groups, which is another way of saying that they have to be uncorrelated with ob-

served group covariates, as all well-behaved error terms are supposed to do. In contrast, the α_i can be seen to control for all unobserved group characteristics that are shared by group members, whether or not they are correlated with the observed covariates. This is a very useful distinction. Econometricians often view fixed effects as random effects which happen to be correlated with the observed covariates.

Third, note that the fixed-effects estimator cannot estimate the effects of group-level variables, or more generally variables that are constant across group members. Otherwise you might think that all we need is the fixed-effects estimator, which is valid under more general conditions. (Incidentally there is a Hausman especification test for random effects which compares the two estimators of the effects for individual-level variables. Just bear in mind that when this test rejects the random specification it doesn't mean that the fixed specification is valid, just that the random is not.)

Fourth, fixed-effect models deal with just two levels, whereas random-effects models can be generalized easily to more than two levels. This can become an important consideration if you have three-level data, for example children, families and communities, and want to study the dependence at all levels.

Fifth, in a random-effects framework we can let any of the coefficients vary from group to group, not just the constant, moving from so-called *random-intercept* models to more interesting *random-slope* models. You can think of a random slope as interacting an individual covariate with unobserved group characteristics. Of particular interest are treatment effects that may vary from group to group. (Or from individual to individual if you have repeated measurements on each person.)

2.4 Between and Within Groups

There's one more way to look at these models. Let us start from the random-effects model and consider the group means, which follow the model

$$\bar{Y}_i = a_i + \bar{x}_i'\beta + \bar{e}_i \tag{3}$$

where we have also averaged the covariates and the error terms for all members of each group. The key fact is that the means follow a linear model with the same regression coefficients β as the individual data.

If the error terms are independent across groups then we can obtain a consistent estimator of β using OLS, or WLS if the number of observations varies by group. (If the a_i are correlated with the x's, however, we have the usual endogeneity problem.) We call this the *between*-groups estimator.

We can also look at deviations from the group means, which follow the model

$$Y_{ij} - \bar{Y}_i = (x_{ij} - \bar{x}_i)'\beta + (e_{ij} - \bar{e}_i)$$

The interesting thing here is that the deviations from the mean also follow a linear model with the same regression coefficients β. The errors are not independent across subjects, but the dependence arises just from subtracting the mean and is easily corrected. We call the resulting estimator the *within*-groups estimator.

It can be shown that the fixed-effects estimator is the same as the within-groups estimator, and that the random-effects estimator is an average or compromise between the between and within estimators, with the precise weight a function of the intra-class correlation.

In the context of multilevel models it is possible to reconcile the fixed and random-effects approaches by considering the group means as additional predictors. Specifically, consider the model

$$Y_{ij} = a_i + \bar{x}_i'\beta_B + (x_{ij} - \bar{x}_i)'\beta_W + e_{ij}$$

where the group mean and the individual's deviation from its group mean appear as predictors. The estimate of β_B, representing the effect of the group average on individual outcomes, coincides with the between-group estimator. The estimate of β_W, representing the effect of an individual's deviation from the group average, coincides with the within-groups or fixed-effects estimator. The random-effects estimator is appropriate only if both coefficients are equal, in which case it is appropriate to average the two estimates.

This more general model can be fitted by OLS to obtain consistent if not fully efficient parameters estimates, but to obtain correct standard errors you would need to correct for clustering at the group level. A much better approach is to fit the model as a random-effects model, in which case maximum likelihood will yield fully-efficient estimates and correct standard errors.

2.5 Examples

We consider two examples, one where the fixed and random-effect approaches lead to similar estimates and one where they differ substantially.

Example 1. Snijders and Bosker (1999) have data for 2287 eighth-grade children in 131 schools in the Netherlands. We are interested in the relationship between verbal IQ and the score in a language test. The table below compares OLS, fixed-effects and random-effects estimators.

```
-------------------------------------------------
Variable     |    ols           re           fe
-------------+-----------------------------------
#1           |
     iq_verb |  2.6538956     2.488094     2.4147722
       _cons |  9.5284841    11.165109    12.358285
-------------+-----------------------------------
sigma_u      |
       _cons |                3.0817186
-------------+-----------------------------------
sigma_e      |
       _cons |                6.4982439
-------------------------------------------------
```

The differences among the three approaches in this particular example are modest. The random-effects model estimates the correlation between the language scores of children in the same school as 0.18. This is equivalent to saying that 18% of the variance in language scores is across schools, and of course 82% is among students in the same school.

The Stata logs also show the regression based on school means, with a coefficient of 3.90, and separate regressions for each school, indicating that the relationship between verbal IQ and language scores varies by school.

Example 2. Wooldridge (2002) has an interesting dataset with the percentage of births classified as low birth weight and the percentage of the population in the AFDC welfare program in each of the 50 states in 1987 and 1990.

We consider models predicting low birth weight from AFDC participation and a dummy for 1990. For simplicity we ignore other controls such as physicians per capita, beds per capita, per capita income, and population (all logged), which turn out not to be needed in the fixed-effects specification. Here are the results:

```
-------------------------------------------------
Variable     |    ols           re           fe
-------------+-----------------------------------
#1           |
         d90 |  .03326679    .14854716    .21247362
     afdcprc |  .26012832   -.01566323   -.16859799
       _cons |  5.6618251    6.6946585    7.2673958
-------------+-----------------------------------
```

```
sigma_u      |
       _cons |               1.1478165
-------------+----------------------------------------
sigma_e      |
       _cons |                .19534447
------------------------------------------------------
```

The OLS estimate suggests that AFDC has a pernicious effect on low birth weight: a higher percentage of the population in AFDC is associated with increased prevalence of low birth weight. The random-effects estimator shows practically no association between AFDC participation and low birth weight. The intra- state correlation is 0.972, indicating that 97% of the variation in low birth weight is across states and only 3% is within states over time. Focusing on intra-state variation, the fixed-effects estimator shows that an increase in the percent of the population in AFDC is associated with a reduction in the percent of low birth-weight births, a much more reasonable result.

My interpretation of these results is that there are unobserved state characteristics (such as poverty) that increase both AFDC participation and the prevalence of low birth weight, inducing a (spurious) positive correlation that masks or reverses the (true) negative effect of AFDC participation on low birth weight. By controlling (implicitly) for all persistent state characteristics, the fixed-effects estimator is able to unmask the negative effect.

The Stata log expands on these analysis using all the controls mentioned above. It also shows how one can reproduce the fixed effects estimate by working with changes between 1987 and 1990 in AFDC participation and in the percent low birth weight, or by working with the original data and introducing a dummy for each state.

3 Binary Data

We now consider extending these ideas to modeling binary data, which pose a few additional challenges. In this section Y_{ij} is a binary outcome which takes only the values 0 and 1.

3.1 Fixed-Effects Logits

In a fixed-effects model we assume that the Y_{ij} have independent Bernoulli distributions with probabilities satisfying

$$\text{logit}(\pi_{ij}) = \alpha_i + x'_{ij}\beta$$

Effectively we have introduced a separate parameter α_i for each group, thus capturing unobserved group characteristics.

Introducing what may be a large number of parameters in a logit model causes the usual practical difficulties and a twist on the theory side. In the usual scenario, where we let the number of groups increase to infinity but not the number of individuals per group, it is not just the estimates of α_i that are not consistent, but the inconsisteny propagates to β as well! This means that there is not point in introducing a separate dummy variable for each group, even if we could.

There is, however, an alternative approach that leads to a consistent estimator of β. We calculate the total number of successes for each group, say $Y_i = \sum_j Y_{ij}$, and look at the distribution of each Y_{ij} given the total Y_i. It turns out that this conditional distribution does not involve the α_i but does depend on β, which can thus be estimated consistently.

In the linear case the dummy and conditioning approaches were equivalent. Here they are not. The conditioning approach requires the existence of a minimal sufficient statistic for the α_i. In logit models the totals have this property. Interestingly, in probit models there is no minimal sufficient statistic for the α_i, which is why there is no such thing as a fixed-effects probit model.

We will skip the details here except to note that conditioning means that groups were all observations are successes (or all are failures) do not contribute to the conditional likelihood. In some situations this can lead to estimating the model in a small subset of the data. This is worrying, but advocates of fixed-effects models argure that those are the only cases with revelant information.

An example may help fix ideas. Suppose one was interested in studying the effect of teenage pregnancy on high school graduation. In order to control for unobserved family characteristics, you decide to use data on sisters and fit a fixed-effects model. Consider families with two sisters. If both graduate from high school, the conditional probability of graduation is one for each sister, and hence the pair is uninformative. If neither graduates the conditional probability of graduation is zero, and thus the pair is also uninformative. It is only when one of the sisters graduates and the other doesn't that we have some information.

So far we have considered variation in the outcome but it turns out that we also need variation in the predictor. If both sisters had a teenage pregnancy the pair provides no information regarding the effect of pregnancy on graduation. The same is true if neither gets pregnant. The only families that contribute information consist of pairs where one sister get pregnant

and the other doesn't, and where one graduates from high school and the other doesn't. The question then becomes whether the one who graduates is the one who didn't get pregnant, an event that can be shown to depend on the parameter of interest and is not affected by unobserved *family* characteristics.

The concern is that very few pairs meet these conditions, and those pairs may be selected on unobserved *individual* characteristics. To see why this is a problem suppose the effect of teenage pregnancy on high school graduation varies with an unobserved individual attribute. The estimated effect can still be interpreted as an average, but the average would be over a selected subset, not the entire population.

3.2 Random-Effects Logits

In a random-effects logit model we postulate the existence of an unobserved individual effect a_i such that *given* a_i the Y_{ij} are independent Bernoulli random variables with probability π_{ij} such that

$$\text{logit}(\pi_{ij}) = a_i + x'_{ij}\beta$$

In other words the *conditional* distribution of the outcomes given the random effects a_i is Bernoulli, with probability following a standard logistic regression model with coefficients a_i and β.

Just as before we treat a_i as an error term and assume a distribution, namely $N(0, \sigma_a^2)$. One difficulty with this model is that the *unconditional* distribution of Y_{ij} involves a logistic-normal integral and does not have a closed form.

This lead several authors to propose approximations, such as marginal quasi-likelihood (MQL) or penalized quasi-likelihood (PQL), but unfortunately these can lead to substantial biases (Rodríguez and Goldman, 1995).

Fortunately it is possible to evaluate the likelihood to a close approximation using *Gaussian quadrature*, a procedure that relies on a weighted sum of conditional probabilities evaluated at selected values of the random effect. These values can be pre-determined or tailored to the data at hand in a procedure known as adaptive Gaussian quadrature, the latest Stata default.

The model can also be formulated in terms of a *latent variable* Y_{ij}^* such that $Y_{ij} = 1$ if and only if $Y_{ij}^* > 0$, by assuming that the latent variable follows a random-effects linear model

$$Y_{ij}^* = a_i + x'_{ij}\beta + e_{ij}$$

where e_{ij} has a standard logistic distribution. The unconditional distribution of Y^* is then logistic-normal and, as noted above, does not have a closed form.

Recall that the variance of the standard logistic is $\pi^2/3$. This plays the role of σ_e^2, the individual variance. We also have the group variance σ_a^2. Using these two we can compute an *intraclass correlation* for the latent variable:

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \pi^2/3}$$

Computing an intra-class correlation for the manifest outcomes is a bit more complicated, as the coefficient turns out to depend on the covariates, see Rodríguez and Elo, 2000) and their `xtrho` command.

3.3 Subject-Specific and Population-Average Models

A common mistake is to believe that all one needs to do with clustered or longitudinal data is to run ordinary regression or logit models and then correct the standard errors for clustering.

This is essentially correct in the linear case, where OLS estimators are consistent but not fully effcient, so all one sacrifices with this approach is a bit of precision. But with logit models, ignoring the random effect introduces a bias in the estimates as well as the standard errors.

To see this point consider a random-effects model, where the expected value of the outcome Y_{ij} given the random effect a_i is

$$E(Y_{ij}|a_i) = \text{logit}^{-1}(a_i + x'_{ij}\beta_{SS})$$

An analyst ignoring the random effects would fit a model where the expected value is assumed to be

$$E(Y_{ij}) = \text{logit}^{-1}(x'_{ij}\beta_{PA})$$

Note that we have been careful to use different notation for the coefficients. We call β_{SS} the subject-specific effect and β_{PA} the population-average effect, because we have effectively averaged over all groups in the population.

In the linear case (just ignore the inverse logit in the above two equations) taking expectation with respect to a_i in the first equation leads to the second, so $\beta_{SS} = \beta_{PA}$ and both approaches estimate the same parameter.

Because of the non-linear nature of the logit function, however, taking expectation in the first equation does not lead to the second. In fact, if the first model is correct the second usually isn't, except approximately.

Typically $|\beta_{PA}| < |\beta_{SS}|$, so population-average effects are smaller in magnitude than subject-specific effects, with the difference increasing with the intra-class correlation.

One could make a case for either model, the main point here is that they differ. From a policy point of view, for example, one could argue that decisions should be based on the average effect. I find this argument more persuasive with longitudinal data, where the averaging is for individuals over time, than with hierarchical data. Suppose you are evaluating a program intended to increase the probability of high school graduation and the model includes a school random effect. Are you interested in the increase in the odds of graduation for students in the school they attend or an hypothetical increase averaged over all the schools in the population?

3.4 Example

Our example comes from the Stata manual and is based on data from the National Longitudinal Survey (NLS) for 4,434 women who were 14-24 in 1968 and were observed between 1 and 12 times each. We are interested in union membership as a function of age, education (grade), and residence, represented by dummy variables for 'not a standard metropolitan area' and the south, plus an interaction between south and time (coded as zero for 1970). We fit ordinary, random-effects, and fixed-effects logit models.

```
------------------------------------------------------
Variable     |    logit        relogit      felogit
-------------+----------------------------------------
#1           |
         age |   .00999311    .00939361    .00797058
       grade |   .04834865    .08678776    .08118077
    not_smsa |  -.22149081   -.25193788    .02103677
       south |  -.71444608   -1.1637691   -1.0073178
     southXt |    .0068356    .02324502    .02634948
       _cons |  -1.8882564   -3.3601312
-------------+----------------------------------------
lnsig2u      |
       _cons |                1.7495341
------------------------------------------------------
```

Compare first the logit and random-effects logit models. We see that, except for age, the subject-specific effects are larger in magnitude than the

population-average effects, as we would expect. For example a woman living in the south in 1970 has 69% lower odds of being a union member than one living elsewhere, everything else being equal. The logit model, however, estimates the average effect as 51% lower odds in 1970. The intraclass correlation measured in a latent scale of propensity to belong to a union is 0.636.

The fixed-effects estimates are in general agreement with the random-effects results except for the indicator for living outside a standard metropolitan area, which changes from -0.252 to $+0.021$. This suggests that the negative association between living outside a SMA and belonging to a union is likely to be spurious, due to persistent unobserved characteristics of women that are associated with both SMA residence and union membership. If we estimate the effect by comparing union membership for the same women when they lived in and outside a SMA we find no association.

Note in closing that we had a total of 26,200 observations on 4,434 women. However, the fixed-effects logit analysis dropped 14,165 observations on 2,744 women because they had no variation over time in union membership.

4 Appendix: Stata Commands

Here's a copy of the do file used to produce the results in this handout.

```
// WWS 509 - Fall 2008 - G. Rodriguez <grodri@princeton.edu>
// Models for Clustered and Longitudinal Data

// Verbal IQ and language scores
use http://data.princeton.edu/wws509/datasets/snijders, clear
reg langpost iq_verb
estimates store ols
xtreg langpost iq_verb, i(schoolnr) mle
estimates store re
xtreg langpost iq_verb, i(schoolnr) fe
estimates store fe
estimates table ols re fe, eq(1 1 1)

// AFDC participation and low birth weight
use http://www.stata.com/data/jwooldridge/eacsap/lowbirth, clear
encode stateabb, gen(stateid)
reg lowbrth d90 afdcprc
estimates store ols
xtreg lowbrth d90 afdcprc, i(stateid) mle
estimates store re
xtreg lowbrth d90 afdcprc, i(stateid) fe
estimates store fe
estimates table ols re fe, eq(1 1 1)

// Union membership
use http://data.princeton.edu/wws509/datasets/union, clear
logit union age grade not_smsa south southXt
estimates store logit
xtlogit union age grade not_smsa south southXt, i(id) re
estimates store relogit
xtlogit union age grade not_smsa south southXt, i(id) fe
estimates store felogit
estimates table logit relogit felogit, eq(1 1 1)
```