Multilevel Models

4. Longitudinal Data. Growth Curve Models

Germán Rodríguez

Princeton University

April 4, 2018

Longitudinal data

MALMUS devotes Chapters 5-7 to models for longitudinal data with emphasis on short panels, and considers four kinds of models

1. Random-effect models, where unobserved heterogeneity at the subject level is represented by random intercepts and slopes
2. Fixed-effect models, where we introduce an additional parameter per subject to focus on within-subject variation
3. Dynamic models, where the response at a given time depends on previous or lagged responses
4. Marginal models, where focus is on population average effects and individual differences are of secondary concern

We will focus on random-effect models for longitudinal data. Many of the issues that arise here are the same as for clustered data, so we will place emphasis on aspects that are unique to panel data. We will then close with a couple of words on dynamic models.

Growth-curve models

We consider a repeated-measurements design where an outcome is measured at different times on the same individuals, leading to a *growth curve* or latent trajectory model.

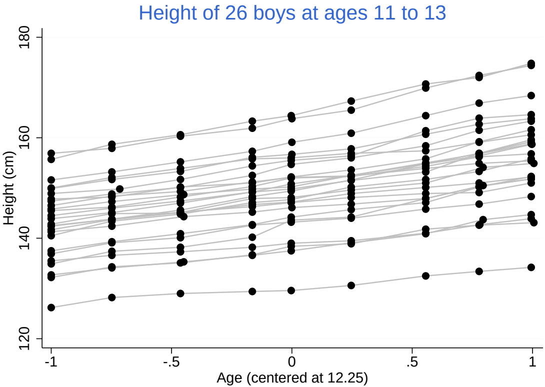
Examples include weight gain during pregnancy, or depression scores by age. The term *latent* trajectory is used because each individual follows his or her own curve over time.

Growth curve models can be fit using standard two-level models where the individual acts as the grouping level, particularly if they are extended to allow for *serial* correlation in the residuals.

If all individuals are measured at exactly the same ages, growth curves can also be modelled using structural equation models (SEM) with exactly the same results for equivalent models.

Height of boys at ages 11 to 13

We illustrate the main ideas using an example in Goldstein (1995), see §6.4 and 6.5, starting on page 91, on the heights of boys measured on nine occasions



The data are available on the course website as `oxboys.dta`, with an analysis using Stata and R at `oxboys.html`

A polynomial growth equation

The basic model used by Goldstein is a fourth-degree polynomial on age, where the constant, linear and quadratic coefficients are random at the child level, so

$$Y_{it} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})x_{it} + (\beta_2 + b_{2i})x_{it}^2 + \beta_3 x_{it}^3 + \beta_4 x_{it}^4 + e_{it}$$

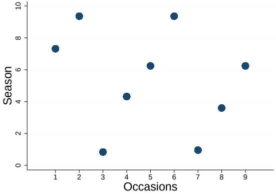
where Y_{it} is height in cm and x_{it} is age of the i-th child at time t, centered around 12 years and 3 months.

The child-level residuals (b_{0i}, b_{1i}, b_{2i}) are assumed to come from a trivariate normal distribution with mean zero and unstructured covariance matrix (with three variances and three correlations), and $e_{it} \sim N(0, \sigma_e^2)$ is the occasion-specific error term.

This is a standard random-coefficient model with the child as the grouping level, so we already know how to fit it. Let's add some bells and whistles.

Seasonality

Observations taken throughout the year may exhibit seasonality. In our dataset the boys were measured in different months of the year, as shown in a plot of `season` by `occasion`



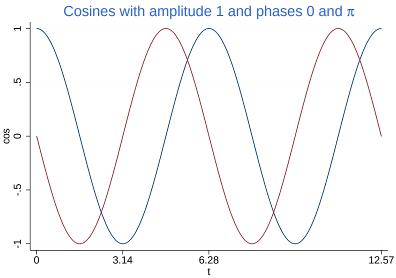
A simple model where a seasonal component has amplitude α and phase ϕ can be written as

$$\alpha \cos(t + \phi) = \alpha_1 \cos(t) - \alpha_2 \sin(t)$$

In this dataset the coefficient of the sine term was very close to zero and was omitted from the model.

Aside on cosines

For those of us who need a refresher, here's a plot of $cos(t)$ for $t \in (0, 4\pi)$ in and out of phase



To compute the cosine term we simply scale `season` to the range $(0, 2\pi)$, calculate

$$\mathrm{sc} = \cos(\pi\,\mathrm{seas}/6)$$

and add the resulting cosine to the fixed part of the model.

The standard model

As this point we are ready to reproduce the results in Table 6.4 in Goldstein (1995, p.93).

Please refer to the website for the code used to run the model in Stata and R. The fixed part of the model has linear, quadratic, cubic and quartic terms on age plus a seasonality term, while the random part lets the intercept and linear and quadratic age terms vary randomly across children.

How would you interpret the coefficient of the seasonality component? How much do you expect a child to grow, on average, between ages 12.25? and 13.25? What's the correlation between the heights of the same child at those two ages? Do you think the model assumptions so far are reasonable?

Serial correlation

With clustered data a random-intercept model assumes an *exchangeable* correlation structure, where any two outcomes have the same correlation, arising from the fact that they share a_i.

With longitudinal data this assumption is suspect because outcomes that are closer in time are likely to be more highly correlated than observations taken further apart.

Fortunately, we can extend the model to allow for *serially correlated* residuals. In particular, we will assume that

$$\text{cov}(e_{it_1}, e_{it_2}) = \sigma_e^2 e^{-\gamma(t_2 - t_1)}$$

which reduces to the variance σ_e^2 when $t_1 = t_2$ and decays exponentially to zero as the gap between the times increases.

Both Stata and R allows for this form of residual correlation, among others.

The full model

The computing logs show how to fit this fourth degree polynomial with seasonality, with the level, gradient and curvature by age varying across children, and residuals that are serially-correlated within each child.

Here are (somewhat abbreviated) results from Stata

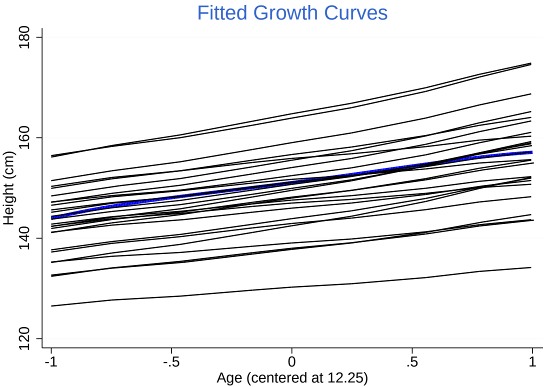
```
Wald chi2(5)    =   502.97
Log likelihood = -305.76024


--------------------------------------
      height |      Coef.   Std. Err.
-------------+------------------------
         age |   6.190767   .3508537
        age2 |    2.16322   .4493732
        age3 |    .386329   .1690328
        age4 |  -1.548466   .4293597
          sc |  -.2360017   .0673323
       _cons |    148.911   1.539373
--------------------------------------
```

```
-----------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.
-----------------------------+-----------------------
id: Unstructured     sd(age) |    1.63716   .2346991
                    sd(age2) |   .7579632    .152763
                   sd(_cons) |   7.840658   1.088743
              corr(age,age2) |   .6869741   .1494221
             corr(age,_cons) |   .6177878   .1243386
            corr(age2,_cons) |   .2489086   .2226974
-----------------------------+------------------------
Residual: Exponential    rho |   .0010001   .0032199
                       sd(e) |    .484354   .0478213
-----------------------------------------------------
```

We will examine these results largely through graphs.

Fitted grow curves

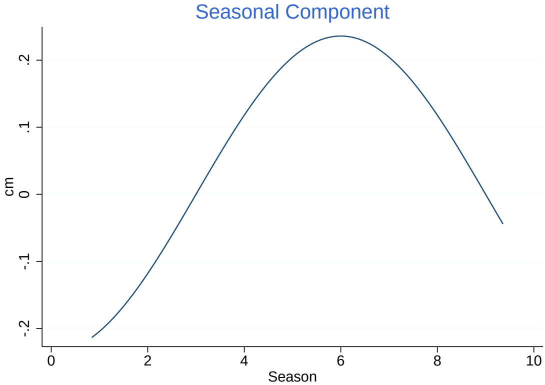
The figure shows the population average curve and the fitted growth curves for each child, using ML to estimate the fixed coefficients and EB for the random coefficients



The curves reflect substantial variation in growth curves across children, with large differences in average height.

Interpreting seasonality

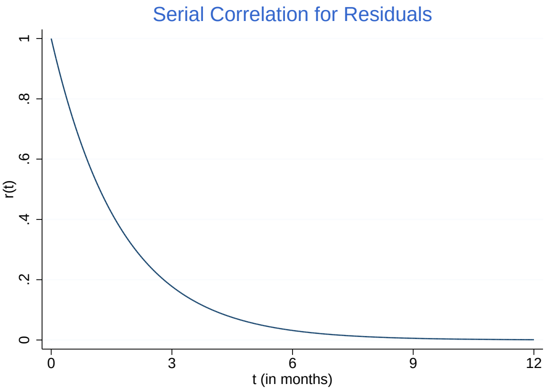
The coefficient of the cosine term or amplitude is estimated at -0.236. We can plot the estimated curve $-0.236\cos(\pi x/6)$ for $x \in (0.84, 9.36)$, the range in the data.



The estimates show that boys grow about half a centimeter more in the summer than in the winter.

Interpreting serial correlation

For residuals with a gap of t the serial correlation is $\rho(t) = e^{-\gamma t}$. Stata reports $\rho(1) = 0.001$ so $\gamma = 6.91$. We plot $\rho(t) = e^{-\gamma t}$ for t in $(0, 1)$, but label the gap in months:



The correlation between residuals is 0.178 after 3 months, and falls to 0.032 after 6 months.

Correlation among outcomes

It is important to understand that the serial correlation we have estimated is just one aspect of the correlation among outcomes in the same child, the part due to correlated residuals.

A larger part of the correlation comes from the latent trajectory, or the fact that measurements on a child on different occasions share the random intercept and slopes for the linear and quadratic terms.

In fact, the correlation between heights measured at ages 11.25 and 11.5, corresponding to the first two occasions, is estimated as 0.996 according to the model. We'll see in a minute how to obtain this result from first principles.

The observed correlation is also 0.996. The easiest way to verify this fact is to change the data to wide format.

Calculating correlations

The outcomes at ages 11.25 and 11.5 for child i involve the random effects $u_i = (a_i, b_i, c_i, e_{i1}, e_{i2})'$.

The variances and covariances of these terms can be extracted from the output and turn out to be

$$V = \begin{bmatrix} 61.476 & & & & \\ 7.930, & 2.689 & & & \\ 1.479, & 0.852, & 0.575 & & \\ 0, & 0, & 0, & 0.235 & \\ 0, & 0, & 0, & 0.042, & 0.235 \end{bmatrix}$$

The random part of the outcomes for the same child at the given ages is a linear combination of u_i with coefficients

$$C = \begin{bmatrix} 1, & -1, & 1, & 1, & 0 \\ 1, & -0.75, & 0.75^2, & 0, & 1 \end{bmatrix}$$

The variance-covariance of u_i is then CVC'.

Testing variances of random coefficients

There is no question that the curves vary by child. The table below shows reductions in deviance starting from the population average model, letting the intercept, slope and curvature be random, and finally allowing for serial correlation of residuals.

Model	log *L*	χ^2	*df*
Fixed coefficients	-819.79		
Random intercept	-463.62	712.33	1
Random slope	-333.26	260.73	2
Random curvature	-306.79	52.93	3
Serial correlation	-305.76	2.06	1

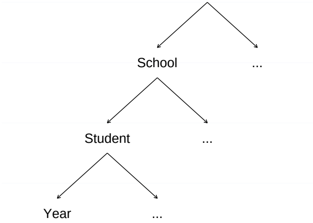
All tests are on a boundary of the parameter space and thus are conservative. All are significant except for serial correlation.

You may want to try using REML estimation to see if that makes a difference in light of the modest sample size.

Three-level Models

The computing logs have an analysis of three-level panel data with 7230 observations on 1721 students in 60 schools.

The outcome of interest is math achievement. The data were collected over six years from first to sixth grade, but not all students have six outcomes, so the panel is not balanced.



The data come from Chapter 4 in the HLM 6 manual and came in three files, which I merged into a single Stata file called `egm.dta`. The analysis may be found in `egm.html`.

A Growth Curve

The models considered in the analysis include

1. a three-level variance components model, which helps introduce intra-level correlations,
2. a growth-curve model where math scores increase linearly with year, with intercept and slopes that vary at the student and school level, and
3. a model where a student's growth curve depends on ethnicity, with different intercept and slopes for whites, blacks and hispanics, and the school average curve depends on the percent of students with low income

We follow Bryk and Raudenbush developing the models level-by-level, which helps determine which cross-level interactions to include.

Dynamic models

Consider a lagged-response model, where the outcomes at previous times are treated as covariates. For example in an autoregressive lag-1 or AR-1 model:

$$Y_{it} = \alpha + \beta x_{it} + \gamma y_{i,t-1} + e_{it}$$

where $e_{it} \sim N(0, \sigma^2)$ with independence across occasions.

This model should only be used if it makes sense to control the effect of the covariates on previous outcomes, or if the effect of the lagged response is itself of interest.

With more than two occasions some outcomes appear on both the right and left-hand sides of the equation. If the process started long before the first occasion and $\gamma < 1$ the process will be stationary.

A related approach controls for baseline conditions.

Dynamic models with random effects

The previous model is often extended by adding a random effect at the individual level to account for correlated residuals

$$Y_{it} = (\alpha + a_i) + \beta X_{it} + \gamma Y_{i,t-1} + e_{it}$$

This model poses special challenges because the lagged outcome is necessarily correlated with the random effect.

Anderson and Hsiao proposed an instrumental variables estimator using a second-order lag.

Arellano and Bond proposed a generalized method of moments estimator using additional instruments.

These approaches are both implemented in Stata, but fall beyond the scope of the course.

The Generalized Linear Mixed Model

All the multilevel models considered in this part of the course are special cases of the GLMM

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}}\,\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times q}{\mathbf{Z}}\,\underset{q\times 1}{\mathbf{u}} + \underset{n\times 1}{\mathbf{e}}$$

where $\mathbf{X}$ is the design matrix for the fixed effects $\boldsymbol{\beta}$, $\mathbf{Z}$ is the design matrix for the random effects $\mathbf{u} \sim N_q(\mathbf{0}, \boldsymbol{\Omega})$ and $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ is the error term. Usually $\boldsymbol{\Omega}$ is block-diagonal by level.

In this model the mean and variance are

$$E(\mathbf{y}) = \mathbf{X}\beta \quad \text{and} \quad \mathrm{var}(\mathbf{y}) = \mathbf{Z}\boldsymbol{\Omega}\mathbf{Z}' + \sigma^2\mathbf{I}$$

Exercise: Write down the model matrices for a two-level random-intercept model with 2 observations per group.

GLMM Estimation

If the parameters in $\boldsymbol{\Omega}$ are known, or more generally conditional on estimates of those parameters, the maximum likelihood estimator of β can be obtained by GLS

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

Inversion of $\mathbf{V}$ takes advantage of its block diagonal structure, so the calculations are reasonably straightforward.

Using this estimator in the multivariate normal likelihood yields a profile likelihood that can then be maximized w.r.t. the parameters in $\boldsymbol{\Omega}$. Goldstein showed how this step can also be done using GLS.

Estimation proceeds by alternating the two steps and usually converges very quickly. Harville showed how the same steps can be adapted to use REML as proposed by Patterson and Thompson. The Longford book has details.