Chapter 5 Longitudinal and Panel Data

Paul E. Johnson

February 28, 2016

These notes include the "Introduction to Part III" pp. 227-245 as well as Chapter 5

1 Glossary

Notation continuing from previous

ξ_{ij}: "xi" total (or combined) individual row-level error term for group i row j,

Longitudinal terminology: "group" is often a single person (j is group of rows) and the rows for each person are differentiated by time (which is i in this book). Economists tend to call these i, and t, whereas Laird-Ware/Bates would call them i j

ζ_j: "zeta" a group level random error, j indexes a grouping variable, $Var(\zeta_j) = \psi$

ϵ_{ij}: "epsilon" $Var(\epsilon_{ij}) = \theta$. Individual row-level error uncorrelated with ζ_j

$Cov(y_{ij}, y_{i'j}|\zeta_j) = 0$. Apart from ζ_j, the observed scores are "conditionally uncorrelated"

2 Introduction to Part III

1. 4 types of models
   a) Random effects: group-dependent intercepts and slopes thought of as random variables
   b) Fixed effects: group-dependent intercepts and slopes thought of as fixed parameters (just more β)
   c) Dynamic models: Lag y as a predictor (with or without random effects)
   d) Marginal models: difficult to understand in general, especially if you are trying to learn what this is from this paragraph. I won't try to clear that up in these notes either.

2. The most powerful, true, fabulous statement ever, p. 229

   > Which of the approaches to longitudinal modeling outlined above is adopted in practice depends largely on the discipline. In biomedical sciences, random-effects and marginal models are most common, whereas random-effects models are popular in most of the social sciences. In economics, fixed-effects models and dynamic models are predominant. Repeated measures or split-plot

analysis of variance (ANOVA) can in some ways be viewed as a fixed-effects model; this approach is used mostly for experimental deisngs in areas such as agriculture and psychology but is increasingly being replaced by random-effects models. Growth-curve modeling is particularly popular in education and psychology.

3. Balance, Strong Balance, Constant Spacing
   a) Balance: equal number of rows within each sub-grouping
   b) Strong Balance: no missing data values (I've not heard this terminology before)
   c) Treating time units as a grouping variable:
   d) Many "out of the box" models only make sense with evenly spaced observations. This may be less true in the future, Prof. Pascal Deboeck at KU has done a lot of work on it.

4. PJ asks: Why is "time" another grouping level? How to think of it?

5. p. 240 Age-period-cohort effects.

   Different ideas about time, how it might be relevant, and not analytically separable.

$$A_{ij} = P_i - C_j$$

   a) C_j: cohort begins in year (birth year, for example)
   b) P_j: period is calendar year
   c) A_{ij}: age of individual i within cohort j

   It is not easy/possible, conceptually or statistically, to separate the effect of "age" from "cohort"

   a) PJ: insert observation.
      i. Cohort effect usually thought of as intercept shift in a regression line
      ii. Effect of aging is coefficient like $\beta_j \times A_{ij}$, if all cohorts experience same effect of aging, then $\beta_j = \beta$
   b) Tough to differentiate the effect of aging from common over-time influences on a group of people who are in an age cohort.
   c) To say "old people are conservative" we have to show that old people who grew up under many different cohorts are all conservative, otherwise the conservatism of today's old people is probably due to a cohort effect that marked a whole generation of people (PJ: get cite for famous Bennington college study claimed a cohort, denied age effect)
   d) See graph p. 240, where time plots for several cohorts are drawn. With many cohorts, can separate (maybe) cohort from aging effects.
   e) In many models described below, there is multicollinearity between a person's age and the period of the data collection.

6. Pooled OLS

   a) Fit one model using all of the observations, ignore grouping, time effects.

   b) RHS Center variables to "make the intercept more interpretable" (p. 241)

```
generate educt = educ - 12
generate yeart = year - 1980
```

      These force the y axis to the "left edge" of the data.

   c) Run a regular regression, but ask for sandwich variance estimates using clusters at the nr level.

```
regress y x1 x2 x3 yeart educt, vce(cluster nr)
```

   d) TODO: find out more about which HCCV approach stata uses, what role does the discrete cluster play in their method?

      i. Claim p. 242 "Furthermore, the sandwich estimator of the standard errors, requested by the vce(cluster nr) option, produces consistent estimate of the standard errors even if the residuals are correlated within subjects and have nonconstant variance."

         TODO: Need to verify this

7. Correlated residuals

   a) Code p. 242-3 (Code below item 10),

   b) `predict res, residuals`

   c) **CAUTION** p. 243 "If running the subsequent commands from a do-file, all the commands from preserve to restore must be run in one block, not one command at a time." !!

```
preserve
keep nr res yeart
reshape wide res, i(nr) j(year)
tabstat res*, statistics(sd) format(%3.2f)
```

      Question: Why is j "year" and not "yeart"

   d) p. 244 "Within-subject correlations over time are sometimes referred to as longitudinal correlations, serial correlations, or autocorrelations, and these term also suggest that correlations may depend on the time interval between occasions."

   e) NOTE: many sub-theories exist and are used to deal with that within-subject over time correlation, data seldom supports inclusion of all sub-theories within a single model.

8. Why bother?

   a) The question:

Because pooled OLS with robust standard errors for clustered data gives consistent estimates of regression coefficients and standard errors (assuming a correctly specified mean structure), the question is, why are there three chapters on longitudinal modeling in this part of the book? (p. 244)

a) Answers

   i. Improved causality analysis. "Indeed, the great advantage of longitudinal data as compared with cross-sectional data is that each subject can serve as his or her own control. Unfortunately, pooled OLS treats longitudinal data as repeated cross-sectional data (where independent samples of subjects are drawn at each occasion) and conflates within- and between-subject comparisons" (p. 244.)

   ii. Omitted variable bias, "unmeasured confounding", time-constant subject specific variables are not included in the model

b) Another limitation of pooled OLS: not consistent "if there are missing data and if missingness depends on observed responses for the same subject, given the covariates. In this common scenario, it becomes necessary ot model the within-subject residual covariance matrix to obtain consistent estimates." (p. 244)

3 Chapter 5 Subject-specific effects and dynamic models

1. Subject specific intercepts. "the intercepts can be viewed as representing the effects of omitted covariates that are constant over time" (p. 247).

2. **Review section 3.4.7**, fixed-effects approach "relaxes the assumption that the covariates are uncorrelated with the subject-specific intercept"(p. 247).

   a) Hm. Check this p. 247 "However, in modern econometrics the subject-specific intercept is usually viewed as random even if the fixed-effects approach is used."

3. Random Intercept model: ζ_j is intercept for data group (person) j

   a) **ASSUMPTIONS.** Recall from section 3.7.4 that "exogeneity assumptions–such as no correlation between covariates and either the random intercept ζ_j or the level-1 residuals ϵ_{ij}–are required for consistent estimation of the parameters in the conventional random intercept model" (p. 248).

   PJ: This drives the agenda in economics. They won't pay attention to an inconsistent estimator, much less one that's unbiased.

   b) **Recall assumptions**, especially $\psi = Var(\zeta_j|X_j)$ and $\theta = Var(\epsilon_{ij}|X_j, \zeta_j)$. "It is assumed that the ζ_j are uncorrelated across subjects, that the ϵ_{ij} are uncorrelated across both subjects and occasions, and that the ζ_j and ϵ_{ij} are uncorrelated" (p. 248).

   c) Can fit with `xtmixed` if we want just the individual specific intercepts (p. 249).

d) Check the ICC

$$ICC : \hat{\rho} = \frac{\hat{\psi}}{\hat{\psi} + \hat{\theta}}$$

e) Interpretation of ICC

i. The importance of group-level intercept

ii. "the within-subject correlation between residuals". If the ICC that meaning in your mind, then it certainly draws attention to the within subject time dependence!

4. RI models accommodating endogenous covariates: group-mean-centering

a) "**As discussed in section 3.7.4**, subject-mean centered covariates are uncorrelated with ζ_j by construction, and the corresponding coefficients can be consistently estimated."

b) Stata code pp. 250-251 shows mean centering, leading to an xtmixed estimate:

```
i. xtmixed y x1dev x2dev x3dev x1mean x2mean x3mean || nr:
      , mle vce(robust)
```

c) "estimated standard errors are valid even if the level-1 errors are heteroskedastic or autocorrelated" (p. 251).

d) "Unfortunately, this approach produces inconsistent estimate for the effects of the time-constant covariates ..., even if they are exogenous. The estimator for the random-intercept variance is also inconsistent" (p. 241).

e) TEST between and within effects are equal (p. 252). I missed the part where this test is supposed to matter to me.

```
test (x1dev = x1mean) (x2dev = x2mean) (x3dev=x3mean)
```

"We conclude that the between and within effects are significantly different from each other, which suggests that one or more of the time-varying covariates are endogenous. This test is numerically identical to the simultaneous test that the coefficients for all the cluster means are zero in a random intercept model that includes the cluster means of the time-varying covariates as well as the noncentered covariates (the approach taken in section 3.7.4). The test is also asymptotically equivalent to the Hausman test discussed in section 3.7.6" (p. 252).

f) Test exogeneity for each of the time-varying covariates by performin separate tests of equal between and within effects. (uses lincom)

i. "For instance, if the random intercept is interpreted as unmeasured ability, then high ability subjects might have higher mean earnings than expected given their observed covariates and be more likely to be union members" (p. 252).

5. Consistent Estimation: Hausman Taylor model. Stata xthtaylor

a) Endogenous/Exogenous, time-varying, time-constant

b) 4 Variable types

	Exogenous	Endogenous
Time-varying	x_{ij}	x_{ij}^{end}
Time-constant	x_j	x_j^{end}

c) Requirements: Must be as many time-varying predictors as endogenous time-constant covariates.

d) Steps that go on inside H-T estimator

   i. Fixed effects estimator of time-varying predictors (section 3.7.2 and 5.4) are consistent.

      A. Get those residuals.

   ii. Regress subject-mean residuals on time constant predictors, using the exogenous covariates as instrumental variables: obtains consistent estimates of time-constant predictors (see display 5.1).

   iii. Use estimates from previous to get residuals which which to estimate of θ and ψ.

   iv. GLS transform of output variables using θ and ψ.

   v. HT estimator uses 3 sets of instruments

      A. deviations from cluster means of time-varying covariates

      B. cluster means of exogenous time-varying covariates

      C. exogenous time-constant covariates.

e) "If the random-intercept model is correctly specified (including the designation of exogenous and endogenous covariates), the Hausman-Taylor method will produce consistent and asymptotically efficient estimators for all model parameters, including the regression coefficients of time-constant covariates and the random-intercept variance"(p. 255).

f) Amemiya and MaCurdy method (`amacurdy` in Strata) is an enhancement for small samples using more instruments.

g) Use `xthtaylor` in Stata (see 13)

h) "The estimates produced by xthtaylor are highly dependent on which covariates are designated as endogenous. Hence, subject-matter considerations regarding endogeneity should be combined with sensitivity analyses to try to ensure sensible estimates. The Hausman-Taylor estimator produces consistent and asymptotically efficient estimates of the coefficients of endogenous time-varying covariates if the model specification is correct. A Hausman test can therefore be used to compare the estimates for the time-varying covariates with their consistent (but possibly inefficient) counerparts from the fixed-effects approach..." (p. 257).

6. Fixed-intercept model (aka **Fixed-Effects** model in econometrics)

a) See section 3.7.2. Can fit by OLS with dummy variables for each subject and omit constant.

b) Also equivalent to a group-mean centered regression, which is how fixed effects models are usually described.

c) This gives consistent estimates of time-varying covariates, "as long as relevant time-varying covariates (confounders) are not omitted from the model."

d) **??** "These estimated within effects are numerically identical to the corresponding ML estimates for the random-intercept model (5.1) with different within and between effects." (p. 257) Need to check that.

e) HOWEVER:

   i. Time-constant covariates CANNOT BE ESTIMATED (multicollinearity).

   ii. The α_j (intercepts) are not consistently estimated if the number of occasions remains fixed and the number of subjects increases. (??)

      This is an incidental parameter problem that is due to the number of parameters α_j increasing as the number of subjects increases.

   iii. "Usually, the subject-specific effects are not of interest, and they can be eliminated by subject-mean centering the responses and covariates (which is what the xtreg command with with the `fe` option does). From this perspective, α_j can also be viewed as random intercepts. Because these intercepts are eliminated, it is not necessary to assume that they are uncorrelated with covariates or that they have a constant variance or a normal distribution. The random intercepts could also be correlated across subjects, for instance, because of clustering in states."

   iv. The mean-centered estimation of coefficients

      A. Start with the original theory, where α_j is the within-group intercept

$$y_{ij} = \alpha_j + \beta_2 x_{2j} + \ldots + \beta_4 x_{4ij} + \ldots + \beta_6 L_{ij} + \beta_7 P_i + \beta_8 E_j + \epsilon_{ij} \quad (1)$$

      B. The within-group means are calculated, and we put them in place of the original variables.

$$\bar{y}_{.j} = \alpha_j + \beta_2 x_{2j} + \ldots + \beta_4 \bar{x}_{4.j} + \ldots + \beta_6 \bar{L}_{.j} + \beta_7 \bar{P} + \beta_8 E_j + \bar{\epsilon}_{.j} \quad (2)$$

   v. Notes

      A. All of the i-subscripted variables are replaced by within group averages, where the i is replaced by a period.

      B. $\bar{\epsilon}.j$ is a theoretical quantity, it is not calculated from data. It is written out as book-keeping

C. x_2 , E_j not changed because it is a group level constant.

vi. Subtract equation (1) from (2), term-by term. The group level deviations about the group means appear in the new regression:

$$y_{ij} - \bar{y}_{.j} = \beta_4 \bar{x}_{4.j} + \ldots + \beta_6 \bar{L}_{.j} + u_{ij} \tag{3}$$

vii. Notes

A. CANCELLATION eliminated α_j as well as x_{2j}.

B. The group-centered variables L and P are redundant, we can only keep one.

C. The new, improved error term has the group mean error removed, so if there were persistent unmeasured group level influences, we abolished them.

viii. "A great advantage of these estimates is that they are not susceptible to bias due to omitted subject-level covariates (level-2 endogeneity)" (p. 258).

ix. Stata xtreg or regress with the built-in differencing operator

7. Hausman Test (section 3.7.6) can be used as a test for endogeneity in the time-varying predictors. Must use FGLS estimate to compare with fixed effects model (DON'T use ML)

a) Study the Hausman output p. 262.

b) H_0 : difference in coefficients not systematic

c) Rejected. Similar to test for the joint hypothesis discussed in 5.3.1. "An advantege of the latter test (based on robust standard errors) ist hat it can be used even if there are heteroskedastic or autocorrelated level-1 errors" (p. 262).

8. Section 5.4.2: ANCOVA concepts. I refuse to write this out.

9. Random-coefficient model. Run xtmixed with random slope. This just assumes away all the problems we've been stressing about through this chapter.

a) p. 266 has very interesting commentary in last paragraph. Concerning random slope coefficients, sometimes they don't make sense in centered data because the effect is always smallest whenever centered x is 0, and depending on how the data is generated, that may be ridiculous. TODO: write out a coherent explanation of that for a lecture.

10. Fixed-coefficient model: Fixed-Effects with group-varying intercepts and slopes.

a) Insert a group-varying slope for the variable L_{ij}, so it is $(\beta_6 + \alpha_{2j})L_{ij}$.

b) If there's no missing data, could first-difference the data to estimate some of the coefficients. (p. 267).

c) Now you see why they introduced the D operator for Stata in the earlier part, when it seemed useless.

11. Lagged-response ("dynamic") models (p. 269)

    a) AR-1 lag model

$$y_{ij} = \beta_1 + \gamma y_{i-1,j} + \beta_2 x_{2j} + \ldots + \epsilon_{ij} \tag{4}$$

        i. If we had only one "group", this would be a time-series exercise.

        ii. We have groups, however. Usually, we don't have as many observations within each group as we might like.

    b) Can fit in stata using the "L." symbol in the regression formula, or by manually creating a lagged version of the variable (see Stata code below, part 14).

12. Lagged-response with subject-specific intercepts: Anderson-Hsaio model

    a) Big specification problem because if there are subject-specific intercepts, then lagged y is certainly correlated with the random intercept. Hence biased estimates.

    b) Suggestion by Anderson and Hsiao (p. 274). Use the second lag of y as an instrumental variable. Use that second lag to predict y-lag-1, and then use the residual in the regression. This method gives consistent estimates of time varying predictors, but does not estimate the time-constant predictors.

    c) Caution about Stata `xtivreg, fd`. Stata uses the second lag of difference, but RHS suggest instead the second lag of y instead. So us Stata `ivregress` function.

13. Arellano-bond model

    a) Use more lags as instruments. Arellano and BOnd (1991) Generalized Method of Moments

        i. Stata's xtabond estimator. (p. 277)

14. Missing data and dropout

    a) Jargon, MAR

    b) Simulation p. 279

    c) xtmixed random intercept model does much better than ordinary pooled OLS estimator.

4 Stata Code, Introduction to Part III

1. Running example: wagepan.dta

    a) use http://www.stata-press.com/data/mlmus3/wagepan

2. Wide and long data formats.

    a) Using reshape, see example of reshape to :

i. p. 231 convert long to wide

b) TODO: find out if this

```
format lwage* %5.3f
```

has a permanent effect on the data, or if it simply alters display. I fear the former.

c) The list function

```
list nr lwage* in 1/5, clean noobs abbreviate(6)
```

d) p. 232 reshape wide to long

3. `xtset` puts a session-long setting in place to tell a data set if it is viewed as a panel, and how observations are grouped.

```
xtset nr year
```

The first layer of grouping is "nr", within nr groups, observations are sorted by year.

4. `xtdescribe`, p. 233

a) `xtdescribe if lwage < .`

Here we find one of the truly odd features of Stata. Missing is "." but inside the guts, it is coded as the largest possible number that the computer is able to record. So displaying cases for which a variable is smaller than "." is same as asking for all observations with finite, smaller than maximal values.

b) Output from `xtdescribe` has a "Pattern" element. I do not understand comments bottom page 233. Need to get output from more example data sets to understand what's going on there.

c) `quietly xtset nr` means don't show the output.

5. p. 234,`xtsum` is a summary. Output format

Var		mean	StdDev	Min	Max	Observations
y	overall	42.5	33	0	100	N=
	between		22	0	100	n=
	within		11	0	100	T=

a) Mean exists only overall here

b) What are "between" and "within" ?

i. between: "between-subject"

ii. within: "within-subject" variability

c) Includes `T` because of balance, otherwise would be `T-bar`

6. p. 235, xttab is a summary for categorical variables. MUST reset xt to disregard time

```
quietly xtset nr
xttab union
```

The interpretation of "between" and "within" still escapes me.

7. p. 235 Graphs: **box plot** of responses over time

   a) basic

   ```
   graph box y, over(year)
   ```

   b) fancier

   ```
   graph box y, over(year) intensity (0) medtype(line) ///
       marker(1, mlabel(nr) mlabsize(vsmall) msym(i) mlabpos
          (0) mlabcol(black)) ///
       ytitle(whatever you want)
   ```

8. Plot observed Trajectories. Trellis plot, p. 236

   a) Basic trellis plot simple, insert ", `by (nr, compact)` at end of a two-way plot.

   ```
   twoway line lwage year, by(nr, compact)
   ```

   b) p. 237. Simple approach of plotting all groups "messy", RHS write out code to randomly select some units for time plots. Note the diagonal x labels in the output

   ```
   format lwage* %9.0g
   sort nr year
   set seed 123123
   generate r = runiform() if year==1980
   * Choose 12 at random as follows
   egen num = rank(r) if r < .
   egen number = mean(num), by(nr)
   twoway line lwage year if number<=12, by(nr, compact)
      ytitle(whatever) xtitle(whatever) xlabel( , angle(45))
   ```

9. Spaghetti plot for those 12 sets, but drawn on top of scatter of all data

   ```
   egen mn_lwage = mean(lwage), by(year)
   sort nr year
   twoway (scatter lwage year, jitter(2) msym(0) msize(tiny)) ///
       (line lwage year if number <= 12, connect(ascending) ///
       lwidth(vthin) lpatt(solid)) ///
       (line mn_lwage year, sort lpatt(longdash)) if lwage >-2,
          ///
       ytitle(whatever) xtitle(whatever) ///
       legend(order( 2 "Individual" 3 "Mean"))
   ```

10. Pooled OLS regression: just run "reg"

    a) Check for residual variance patterns.

b) `predict res, residuals`

c) **CAUTION** p. 243 "If running the subsequent commands from a do-file, all the commands from preserve to restore must be run in one block, not one command at a time." !!

```
preserve
keep nr res yeart
reshape wide res, i(nr) j(year)
tabstat res*, statistics(sd) format(%3.2f)
correlate res*, wrap
```

Question: Why is j "year" and not "yeart"

11. HYPO test: TEST between and within effects are equal (p. 252).

```
test (x1dev = x1mean) (x2dev = x2mean) (x3dev=x3mean)
```

12. p. 252. Test exogeneity for each of the time-varying covariates by performin separate tests of equal between and within effects. (uses lincom)

```
lincom x1dev-x1mean
```

Where x1dev is the deviation within group and x1mean is mean within group.

13. p. 256. Hausman-Taylor

```
xtset nr
xthtaylor y x1 x2 x3 x4 x5, endog (x1 x2)
```

Stata "keeps track of whether covariates are time varying or time constant". Note the predictors are not specified in group mean-centered form, the procedure does that inside. Also, the grouping variable is not declared inside the xthtaylor code, it is obtained from environment xtset.

14. LAG: create lag-1 response

```
by nr (yeart), sort: generate lag1 = lwage[_n-1]
```

NOTE: creates a missing value in the first row within each group. `_n` is a row counter in Stata

a) Insert "lag1" as predictor

```
regress lwage lag1 x2 x3...
```

b) Or use Stata symbol "L" in the regression, save the trouble of creating the variable explicitly

```
regress lwage L.lwage x2 x3...
```

15. Anderson-Hsaio lagged y with subject-specific intercepts.

```
xtset nr yeart
ivretgress 2sls D.lwage D.(union married) (LD.lwage = L2.1wage
   )
```

Gives instrumental variables estimates via 2SLS.

16. Arellano-Bond p. 277

```
xtset nr yeart
xtabond lwage union married exper, lags(1) twostep noconstant vce(
   robust)
```