STATISTICAL HORIZONS

Longitudinal Data Analysis Using R

Stephen Vaisey, Ph.D.

Upcoming Seminar:

August 1-2, 2019, Philadelphia, Pennsylvania

Longitudinal Data Analysis Using Stata

This handbook, which was prepared by Paul Allison in June 2018, closely parallels the slides for Stephen Vaisey's course on **Longitudinal Data Analysis Using R**.

Stata data sets for the examples and exercises can be downloaded at

StatisticalHorizons.com/resources/data-sets

STATISTICAL HORIZONS
www.StatisticalHorizons.com

Table of Contents

Outline

1. Opportunities and challenges of panel data.
    a. Data requirements
    b. Control for unobservables
    c. Determining causal order
    d. Problem of dependence
    e. Software considerations
2. Linear models
    a. Robust standard errors
    b. Generalized least squares with ML
    c. Random effects models
    d. Fixed effects models
    e. Between-within models
3. Logistic regression models
    a. Robust standard errors
    b. GEE
    c. Subject-specific vs. population averaged methods
    d. Random effects models
    e. Fixed effects models
    f. Between-within models
4. Count data models
    a. Poisson models
    b. Negative binomial models
5. Linear structural equation models
    a. Fixed and random effects in the SEM context
    b. Models for reciprocal causation with lagged effects

Panel Data

Data in which variables are measured at multiple points in time for the same individuals.

Response variable y_{it} with $t = 1, 2, \ldots, T$

Vector of predictor variables x_{it}.

Some of these may vary with time, others may not.

Assume, for now, that time points are the same for everyone in the sample. (For some methods that assumption is not essential).

Why Are Panel Data Desirable?

In *Econometric Analysis of Panel Data* (2008), Baltagi lists six potential benefits of panel data:

1. Ability to control for individual heterogeneity.
2. More informative data: more variability, less collinearity, more degrees of freedom and more efficiency.
3. Better ability to study the dynamics of adjustment. For example, a cross-sectional survey can tell you what proportion of people are unemployed, but a panel study can tell you the distribution of spells of unemployment.
4. Ability to identify and measure effects that are not detectable in pure cross-sections or pure time series. For example, if you want to know whether union membership increases or decreases wages, you can best answer this by observing what happens when workers move from union to non-union jobs, and vice versa.
5. Ability to construct and test more complicated behavioral models than with purely cross-section or time-series data. For example, distributed lag models may require fewer restrictions with panel data than with pure time-series data.
6. Avoidance of aggregation bias. A consequence of the fact that most panel data are micro-level data.

My List

1. Ability to control for unobservables.

    Accomplished by fixed effects methods.

2. Ability to investigate causal ordering: Does *y* cause *x* or does *x* cause *y*?

    Accomplished by simultaneous estimation of models with lagged predictors.

    Methods for combining fixed effects with cross-lagged models have only recently been developed and not often used (outside of economics)

3. Ability to study the effect of a "treatment" on the trajectory of an outcome (or, equivalently, the change in a treatment effect over time).

Problems with Panel Data

1. Attrition and missing data.

2. Statistical dependence among multiple observations from the same individual.

- Repeated observations on the same individual are likely to be positively correlated. Individuals tend to be persistently high or persistently low.
- But conventional statistical methods assume that observations are independent.
- Consequently, estimated standard errors tend to be too low, leading to test statistics that are too high and *p*-values that are too low.
- Also, conventional parameter estimates may be statistically inefficient (true standard errors are higher than necessary).

- Many different methods to correct for dependence:
  - Robust standard errors
  - Generalized least squares
  - Generalized estimating equations (GEE)
  - Random effects (mixed) models
  - Fixed-effects models
- Many of these methods can also be used for clustered data that are not longitudinal, e.g., students within classrooms, people within neighborhoods.

Software

I'll be using Stata 15, with a focus on the **`xt`** and **`me`** commands.

These commands require that the data be organized in the "long form" so that there is one record for each individual at each time point, with an ID number that is the same for all records for the same individual, and a variable that indicates which time point the record comes from. The "wide form" has one record per individual.

All of the methods described here can also be implemented in SAS.

Linear Models for Quantitative Response

Notation:

y_{it} is the value of the response variable for individual *i* at time *t*.

z_i is a column vector of variables that describe individuals but do not vary over time

x_{it} is a column vector of variables that vary both over individuals and over time

Basic model:

$$y_{it} = \mu_t + \beta x_{it} + \gamma z_i + \varepsilon_{it}\,, \qquad i=1,\ldots,n\,;\ t=1,\ldots,T$$

where ε is a random error term with mean 0 and constant variance, assumed to be uncorrelated with x and z.

β and γ are row vectors of coefficients.

No lags, different intercepts at each time point, coefficients the same at all time points.

Consider OLS (ordinary least squares) estimation.

- Coefficients will be unbiased but not efficient. An efficient estimator is one whose true standard error is as small as possible, i.e., minimal variability across repeated samples.
- Estimated standard errors will be too low because corr($\varepsilon_{it}, \varepsilon_{it'}$) ≠ 0

Example:

581 children interviewed in 1990, 1992, and 1994 as part of the National Longitudinal Survey of Youth (NLSY).

Time-varying variables:

ANTI antisocial behavior, measured with a scale ranging from 0 to 6.

SELF self-esteem, measured with a scale ranging from 6 to 24.

POV poverty status of family, coded 1 for in poverty, otherwise 0.

Time-invariant variables:

BLACK 1 if child is black, otherwise 0

HISPANIC 1 if child is Hispanic, otherwise 0

CHILDAGE child's age in 1990

MARRIED 1 if mother was currently married in 1990, otherwise 0

GENDER 1 if female, 0 if male

MOMAGE mother's age at birth of child

MOMWORK 1 if mother was employed in 1990, otherwise 0

Original data set `nlsy.dta` has 581 records, one for each child, with different names for the variables at each time point, e.g., ANTI90, ANTI92 and ANTI94.

Before converting from the wide form to the long form, let's look at the over-time correlations for the dependent variable.

```
use c:\data\nlsy.dta, clear
corr anti*
```

```
(obs=581)
             |   anti90   anti92   anti94
-------------+---------------------------
      anti90 |   1.0000
      anti92 |   0.6380   1.0000
      anti94 |   0.5447   0.6008   1.0000
```

Note that the 4-year lag correlation is smaller than the two 2-year lag correlations.

Using the **reshape** command, we now convert the data into the long form, a set of 1743 records, one for each child in each year:

```
use c:\data\nlsy.dta, clear
gen id = _n
reshape long anti self pov, i(id) j(year)
save persyr3, replace
```

```
Data                               wide   ->   long
---------------------------------------------------------
Number of obs.                      581   ->    1743
Number of variables                  17   ->      12
j variable (3 values)                     ->   year
xij variables:
                   anti90 anti92 anti94   ->   anti
                   self90 self92 self94   ->   self
                      pov90 pov92 pov94   ->   pov
---------------------------------------------------------
```

Note:

The time-invariant variables are repeated across the multiple records for each child.

The variable **id** has a unique ID number for each child.

The variable **year** has values of 90, 92 or 94.

Now we'll do OLS regression, with no correction for dependence

```
reg anti self pov black hispanic childage married
   gender momage momwork i.year
```

```
      Source |       SS       df       MS              Number of obs =    1743
-------------+------------------------------           F( 11,  1731) =   15.16
       Model |   380.85789    11  34.6234446           Prob > F      =  0.0000
    Residual |  3952.25743  1731  2.28322208           R-squared     =  0.0879
-------------+------------------------------           Adj R-squared =  0.0821
       Total |  4333.11532  1742  2.48743704           Root MSE      =   1.511

------------------------------------------------------------------------------
        anti |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        self |  -.0741425   .0109632    -6.76   0.000     -.095645   -.0526401
         pov |   .4354025   .0855275     5.09   0.000     .2676544    .6031505
       black |   .1678622   .0881839     1.90   0.057    -.0050959    .3408204
    hispanic |  -.2483772   .0948717    -2.62   0.009    -.4344523   -.0623021
    childage |    .087056   .0622121     1.40   0.162    -.0349628    .2090747
     married |  -.0888875    .087227    -1.02   0.308    -.2599689     .082194
      gender |  -.4950259   .0728886    -6.79   0.000     -.637985   -.3520668
      momage |  -.0166933   .0173463    -0.96   0.336    -.0507153    .0173287
     momwork |   .2120961   .0800071     2.65   0.008     .0551754    .3690168
       year  |
         92  |   .0521538   .0887138     0.59   0.557    -.1218437    .2261512
         94  |   .2255775   .0888639     2.54   0.011     .0512856    .3998694
       _cons |   2.675312   .7689554     3.48   0.001     1.167132    4.183491
```

Problems:

Although the coefficients are unbiased, they are not efficient (true standard errors are larger than necessary).

More important, reported standard errors and *p*-values are probably too low

Solution 1: Robust standard errors

Robust standard errors are standard error estimates that correct for dependence among the repeated observations. Also known as Huber-White standard errors, sandwich estimates, or empirical standard errors.

For OLS linear models, conventional standard errors are obtained by first calculating the estimated covariance matrix of the coefficient estimates:

$$s^2(\mathbf{X}'\mathbf{X})^{-1}$$

where s^2 is the residual variance and **X** is a matrix of dimension $Tn \times K$ (n is the number of individuals, T is the number ot time periods, and K is the number of coefficients). Standard errors are obtained by taking the square roots of the main diagonal elements of this matrix.

The formula for the robust covariance estimator is

$$\hat{\mathbf{V}} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_i \mathbf{X}_i'\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)'\mathbf{X}_i\right)(\mathbf{X}'\mathbf{X})^{-1}$$

where $\mathbf{X}_i$ is a T x K matrix of covariate values for individual i and $\mathbf{y}_i$ is a T x 1 vector of y values for individual i. The robust standard errors are the square roots of the main diagonal elements of $\hat{\mathbf{V}}$.

In Stata, this method can be implemented with most regression commands using the `vce` option:

```
reg anti self pov black hispanic childage married
   momage gender momwork i.year, vce(cluster id)
```

```
Linear regression                                   Number of obs =    1743
                                                    F( 11,   580) =    8.99
                                                    Prob > F      =  0.0000
```

```
                                                R-squared     =  0.0879
                                                Root MSE      =   1.511

                                (Std. Err. adjusted for 581 clusters in id)
            |               Robust
       anti |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
       self |  -.0741425   .0133707    -5.55   0.000    -.1004034   -.0478816
        pov |   .4354025   .1093637     3.98   0.000     .2206054    .6501995
      black |   .1678622   .1309221     1.28   0.200    -.0892769    .4250014
   hispanic |  -.2483772   .1341785    -1.85   0.065    -.5119122    .0151578
   childage |    .087056   .0939055     0.93   0.354    -.0973804    .2714923
    married |  -.0888875   .1336839    -0.66   0.506    -.3514509     .173676
     momage |  -.0166933   .0241047    -0.69   0.489    -.0640364    .0306498
     gender |  -.4950259   .1057334    -4.68   0.000    -.7026929   -.2873589
    momwork |   .2120961   .1189761     1.78   0.075    -.0215803    .4457725
       year |
        92  |   .0521538   .0540096     0.97   0.335    -.0539244     .158232
        94  |   .2255775   .0641766     3.51   0.000     .0995306    .3516245
      _cons |   2.675312   1.138426     2.35   0.019     .4393717    4.911252
```

Although coefficients are the same, almost all the standard errors are larger. This makes a crucial difference for MOMWORK, BLACK and HISPANIC.

Notes:

- Robust standard errors may be *smaller* than conventional standard errors.
- You generally see a bigger increase in the standard errors for time-invariant variables than for time-varying variables. Standard errors for time itself often decrease.
- Robust SEs are also robust to heteroscedasticity and non-normality.
- In small samples, robust standard errors may be inaccurate and have low power. For reasonably accurate results, you need *at least* 20 clusters if they are approximately balanced, 50 if they are unbalanced. See Cameron & Miller (2015) *Journal of Human Resources*

Solution 2: Generalized Least Squares (GLS) with Maximum Likelihood.

The attraction of this method is that it, in addition to getting the standard errors right, it produces efficient estimates of the coefficients (i.e., true standard errors will be optimally small). It does this by taking the over-time correlations into account when producing the coefficient estimates.

Conventional least squares estimates are given by the matrix formula

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

GLS estimates are obtained by

$$(\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$$

where $\hat{\mathbf{\Omega}}$ is an estimate of the covariance matrix for the error terms. For panel data, this will typically be a "block-diagonal" matrix. For example, if the sample consists of three people with two observations each, the covariance matrix will look like

$$\hat{\mathbf{\Omega}} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & 0 & 0 & 0 & 0 \\ \hat{\sigma}_{12} & \hat{\sigma}_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \hat{\sigma}_{11} & \hat{\sigma}_{12} & 0 & 0 \\ 0 & 0 & \hat{\sigma}_{12} & \hat{\sigma}_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & \hat{\sigma}_{11} & \hat{\sigma}_{12} \\ 0 & 0 & 0 & 0 & \hat{\sigma}_{12} & \hat{\sigma}_{22} \end{bmatrix}$$

where $\hat{\sigma}_{11}$ is an estimate of var(ε_{i1}), $\hat{\sigma}_{22}$ is an estimate of *var*(ε_{i2}), and $\hat{\sigma}_{12}$ is an estimate of cov(ε_{i1}, ε_{i2}). We assume that these variances and covariances are the same across individuals.

There are many different ways to estimate these variances and covariances. I used to focus on the method of generalized estimating equations (GEE), as implemented with the **xtgee** command. We will use this method later for logistic regression. For linear models, I now prefer maximum likelihood, implemented with the **mixed** command:

```
mixed anti self pov black hispanic childage married gender
    momage momwork i.year || id:, noconstant
    residuals(unstructured,t(year)) stddeviations
```

- **noconstant** says "don't fit a random intercepts model" (see the next section).
- **residuals** says "estimate variances and covariances for the error terms."
- **unstructured** says "don't impose any structure on variances and covariances."
- **t(year)** sets the time dimension.
- **stddeviations** says "report standard deviations and correlations instead of variances and covariances."

```
Mixed-effects ML regression                     Number of obs     =      1,743
Group variable: id                              Number of groups  =        581
                                                Obs per group:
                                                              min =          3
                                                              avg =        3.0
                                                              max =          3

                                                Wald chi2(11)     =     105.93
Log likelihood = -2910.4053                     Prob > chi2       =     0.0000
------------------------------------------------------------------------------
        anti |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        self |  -.0597951   .0093479    -6.40   0.000    -.0781166   -.0414736
         pov |   .2739148   .0797349     3.44   0.001     .1176373    .4301924
       black |   .2221163   .1234505     1.80   0.072    -.0198422    .4640748
    hispanic |  -.2369447   .1358772    -1.74   0.081    -.5032592    .0293697
    childage |   .0607793   .0895205     0.68   0.497    -.1146778    .2362363
     married |  -.0376591   .1244933    -0.30   0.762    -.2816615    .2063433
      gender |  -.4967857   .1046405    -4.75   0.000    -.7018773   -.2916942
      momage |  -.0150424   .0248672    -0.60   0.545    -.0637811    .0336964
     momwork |   .2661671   .1127786     2.36   0.018      .045125    .4872091
```

```
             |
        year |
         92  |   .0468588   .0532737     0.88   0.379    -.0575558    .1512734
         94  |   .2155011   .0628523     3.43   0.001     .0923129    .3386894
             |
       _cons |   2.590622   1.075436     2.41   0.016     .4828067    4.698437
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id:                  (empty) |
-----------------------------+------------------------------------------------
Residual: Unstructured       |
                     sd(e90) |   1.402862   .0413735      1.324071    1.486342
                     sd(e92) |    1.47728    .043512      1.394413    1.565072
                     sd(e94) |   1.635629   .0482859      1.543676    1.733059
                corr(e90,e92) |   .6045961   .0264302       .550231    .6538571
                corr(e90,e94) |   .5151491   .0308114      .4522374      .57296
                corr(e92,e94) |   .5836673   .0274119      .5273888    .6348463
------------------------------------------------------------------------------
LR test vs. linear model: chi2(5) = 552.57                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 coefficient for POV is noticeably smaller, although still highly significant. The coefficient for MOMWORK is somewhat larger, and the *p*-value is again well below .05.

The over-time correlations for the errors are all quite large. At the bottom we get a likelihood ratio chi-square test of the null hypothesis that (a) the three correlations are 0, and (b) the three standard deviations are the same. This is clearly rejected.

With five or fewer time points, the unstructured model is usually the best way to go. With many time points the number of unique correlations will get large: $T(T\text{-}1)/2$. And unless the sample is also large, estimates of all these parameters may be unstable.

In that case, consider restricted models. Let ρ_{ts} be the correlation between ε_{it} and ε_{is}, i.e., errors for the same individual at time *t* and time *s*. Here are some possible structures for longitudinal data:

TYPE	Description	Formula
EXCH	Exchangeable, i.e., equal correlations	$\rho_{ts} = \rho$
AR#	Autoregressive of order #	$\varepsilon_{it} = \sum_{j=1}^{\#} \theta_j \varepsilon_{it-j} + \nu_{it}$
TOEPLITZ#	Correlation depends on distance between *t* and *s*	$\rho_{ts} = \rho_{\|t-s\|}$ when |*t*-*s*| ≤ #, otherwise $\rho_{ts} = 0$
EXPONENTIAL	A generalization of AR1 to allow unequal gaps	$\rho_{ts} = \rho^{\|t-s\|}$

Results will often be robust to choice of correlation structure, but sometimes it can make a big difference. Unfortunately, all four of these structures presume that the error variances are constant over time. No way to relax that.

The maximum order of AR and Toeplitz is one less than the number of time points. That’s the default for Toeplitz (good) but the default for AR is 1 (bad). With AR1 the correlation usually goes down too rapidly with the distance between measurements.

Let’s try the exchangeable structure with the NLSY data:

```
mixed anti self pov black hispanic childage married
   gender momage momwork i.year || id:, nocon
   res(exch) stddev
```

Notice that you don’t need the **t(year)** option with this correlation structure.