

# Longitudinal Data Analysis Using SAS

Paul D. Allison, Ph.D.

*Upcoming Seminar:*  
May 28-29, 2020, Philadelphia

## 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 via ML
  - c. Random effects models
  - d. Fixed effects models
  - e. Between-Within models
3. Logistic regression models
  - a. Robust standard errors
  - b. Generalized estimating equations
  - 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 framework
  - 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  $t = 1, 2, \dots, T$

Vector of predictor variables  $x_{it}$ .

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

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

## **Why are panel data desirable?**

In *Econometric Analysis of Panel Data* (2005), 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 unobserved variables.

Accomplished by fixed effects methods.

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

Accomplished by estimating bi-directional models with lagged predictors (“cross-lagged models”).

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
- 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 SAS® 9.4. The following procedures will be covered: GLM, SURVEYREG, GENMOD, MIXED, LOGISTIC, SURVEYLOGISTIC, GLIMMIX, CALIS, PANEL

Stata is also an excellent package for panel data analysis, especially the **xt** and **me** commands.

Most software for panel data requires that the data are 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 that come from the same individual, and a variable that indicates which time point the record comes from. The “wide form” (also known as flat data) has one record per person.

# Linear Models for Quantitative Response Variables

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:

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

where  $\varepsilon$  is a random error term with mean 0 and constant variance, uncorrelated with  $x$  and  $z$ .

$\beta$  and  $\gamma$  are row vectors of coefficients.

No lags, different intercepts at each time point, coefficients are the same across times.

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—minimal variability across repeated samples.
- Estimated standard errors will be too low because  $\text{corr}(\varepsilon_{it}, \varepsilon_{it'}) \neq 0$

## Example:

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

### Time-varying variables (measured at each of the three time points):

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

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

POV      poverty status of family, coded 1 for family 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 has 581 records, one for each child (wide form), with different names for the variables at each time point, e.g., ANTI1, ANTI2 and ANTI3.

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

```
PROC CORR DATA=my.nlsy NOPROB;  
VAR anti1 anti2 anti3; RUN;
```

**Pearson Correlation Coefficients, N = 581**

	<b>anti1</b>	<b>anti2</b>	<b>anti3</b>
<b>anti1</b>	1.00000	0.63796	0.54466
child antisocial behavior in 1990			
<b>anti2</b>	0.63796	1.00000	0.60080
child antisocial behavior in 1992			
<b>anti3</b>	0.54466	0.60080	1.000
child antisocial behavior in 1994			

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

The following program converted the data into a set of 1743 records, one for each child in each year:

```
DATA my.nlsy3;  
SET my.nlsy;  
time=1;  
anti=anti1;  
self=self1;  
pov=pov1;  
OUTPUT;  
time=2;  
anti=anti2;  
self=self2;
```



```

    pov=pov2;
    OUTPUT;
time=3;
    anti=anti3;
    self=self3;
    pov=pov3;
    OUTPUT;
DROP anti1-anti3 self1-self3 pov1-pov3;
RUN;

```

My convention: In SAS programs, any word in upper case is part of the SAS language; any word in lower case is a data set name or variable name specific to the example. SAS itself doesn't distinguish upper and lower case (with a few exceptions).

Note:

- The time-invariant variables are replicated across the multiple records for each child.
- The variable TIME has values of 1, 2 or 3.

Here's how to accomplish the same thing with the MAKELONG macro, available at

[www.sascommunity.org/wiki/File:Transpose\\_Macros\\_MAKEWIDE\\_and\\_MAKELONG.sas](http://www.sascommunity.org/wiki/File:Transpose_Macros_MAKEWIDE_and_MAKELONG.sas)

```

%MAKELONG(DATA=my.nlsy, OUT=my.nlsy3, ID=id, COPY=black
    hispanic chldage married gender momage momwork,
    ROOT=anti self pov, MEASUREMENT=time)

```

PROC PANEL can also convert from wide ("flat") to long, but the variables names must be in the form of ANTI\_1, ANTI\_2, etc., and you have to fit a model as well. Also PANEL does not like the ID variable to be called ID.

Here's the program for OLS regression, with no correction for dependence

```
PROC GLM DATA=my.nlsy3;
CLASS time;
MODEL anti=self pov black hispanic chldage
      married gender momage momwork time /SOLUTION;
RUN;
```

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		2.900889443 B	0.77054099	3.76	0.0002
self		-0.074142506	0.01096317	-6.76	<.0001
pov		0.435402473	0.08552747	5.09	<.0001
black		0.167862234	0.08818389	1.90	0.0571
hispanic		-0.248377211	0.09487165	-2.62	0.0089
chldage		0.087055958	0.06221206	1.40	0.1619
married		-0.088887477	0.08722703	-1.02	0.3083
gender		-0.495025904	0.07288865	-6.79	<.0001
momage		-0.016693309	0.01734634	-0.96	0.3360
momwork		0.212096097	0.08000707	2.65	0.0081
time	1	-0.225577516 B	0.08886389	-2.54	0.0112
time	2	-0.173423729 B	0.08870053	-1.96	0.0507
time	3	0.000000000 B	.	.	.

Problem:

Although the coefficients are unbiased, they are not efficient (true standard errors are larger than necessary), and 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  $\mathbf{X}$  is a matrix of the predictor variables with dimension  $Tn \times K$ . ( $n$  is the number of individuals,  $T$  is the number of 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

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

where  $\mathbf{X}_i$  is a  $T \times K$  matrix of covariate values for individual  $i$  and  $\mathbf{y}_i$  is a  $T \times 1$  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 SAS, this method can be implemented with PROC GENMOD and the REPEATED statement:

```
PROC GENMOD DATA=my.nlsy3;  
  CLASS id time;
```

```

MODEL anti=self pov black hispanic childage
      married gender momage momwork time;
REPEATED SUBJECT=id;
RUN;

```

Note: The ID variable must be declared in a CLASS statement.

Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept	2.9009	1.1331	0.6801	5.1217	2.56	0.0105
self	-0.0741	0.0133	-0.1002	-0.0480	-5.57	<.0001
pov	0.4354	0.1089	0.2219	0.6489	4.00	<.0001
black	0.1679	0.1304	-0.0877	0.4234	1.29	0.1980
hispanic	-0.2484	0.1336	-0.5103	0.0136	-1.86	0.0631
childage	0.0871	0.0935	-0.0963	0.2704	0.93	0.3520
married	-0.0889	0.1331	-0.3498	0.1721	-0.67	0.5044
gender	-0.4950	0.1053	-0.7014	-0.2886	-4.70	<.0001
momage	-0.0167	0.0240	-0.0637	0.0304	-0.70	0.4868
momwork	0.2121	0.1185	-0.0202	0.4443	1.79	0.0735
time 1	-0.2256	0.0639	-0.3509	-0.1003	-3.53	0.0004
time 2	-0.1734	0.0595	-0.2900	-0.0568	-2.92	0.0036
time 3	0.0000	0.0000	0.0000	0.0000	.	.

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

An alternative to GENMOD is PROC SURVEYREG with the CLUSTER statement:

```

PROC SURVEYREG DATA=my.nlsy3;
  CLASS time;
  MODEL anti=self pov black hispanic chldage
    married gender momage momwork time / SOLUTION;
  CLUSTER id; RUN;

```

This uses a slightly different method to calculate the robust standard errors, but results are usually almost identical.

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	2.9008894	1.13764794	2.55	0.0110
self	-0.0741425	0.01337073	-5.55	<.0001
pov	0.4354025	0.10936365	3.98	<.0001
black	0.1678622	0.13092208	1.28	0.2003
hispanic	-0.2483772	0.13417850	-1.85	0.0647
chldage	0.0870560	0.09390554	0.93	0.3543
married	-0.0888875	0.13368386	-0.66	0.5064
gender	-0.4950259	0.10573338	-4.68	<.0001
momage	-0.0166933	0.02410468	-0.69	0.4889
momwork	0.2120961	0.11897605	1.78	0.0752
time 1	-0.2255775	0.06417664	-3.51	0.0005
time 2	-0.1734237	0.05972255	-2.90	0.0038
time 3	0.0000000	0.00000000	.	.

Notes:

- Robust standard errors may sometimes 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.

- For small samples, robust standard errors may be inaccurate and have low power. You need *at least* 20 clusters if they are approximately balanced (equal size), 50 if they are unbalanced. See Cameron & Miller (2015) *Journal of Human Resources*
- Robust standard errors are also robust to heteroscedasticity.

## **Solution 2: Generalized Least Squares (GLS) via 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 there are three people with two observations each, the covariance matrix will look like

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

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 PROC GENMOD. We will use this method later for logistic regression. For linear models, I now prefer maximum likelihood, implemented with PROC MIXED:

```
PROC MIXED DATA=my.nlsy3;
  CLASS time;
  MODEL anti=self pov black hispanic chldage
    married gender momage momwork time /SOLUTION;
  REPEATED time / SUBJECT=id TYPE=UNR; RUN;
```

The REPEATED statement implements GLS—otherwise just OLS. TIME has to be a CLASS variable. UNR requests an unstructured correlation matrix, reporting correlations rather than covariances. If the data are sorted by ID (as is often the case), ID does not have to be a CLASS variable—otherwise it does.

#### Model Information

<b>Data Set</b>	MY.NLSY3
<b>Dependent Variable</b>	anti
<b>Covariance Structure</b>	Unstructured using Correlations
<b>Subject Effect</b>	id
<b>Estimation Method</b>	REML
<b>Residual Variance Method</b>	None
<b>Fixed Effects SE Method</b>	Model-Based
<b>Degrees of Freedom Method</b>	Between-Within

#### Covariance Parameter Estimates

<b>Cov Parm</b>	<b>Subject</b>	<b>Estimate</b>
<b>Var(1)</b>	id	1.9929
<b>Var(2)</b>	id	2.2076
<b>Var(3)</b>	id	2.7015

### Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Corr(2,1)	id	0.6081
Corr(3,1)	id	0.5190
Corr(3,2)	id	0.5868

### Fit Statistics

-2 Res Log Likelihood	5864.5
AIC (Smaller is Better)	5876.5
AICC (Smaller is Better)	5876.6
BIC (Smaller is Better)	5902.7

### Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	556.34	<.0001

### Solution for Fixed Effects

Effect	time	Estimate	Standard Error	DF	t Value	Pr >  t
Intercept		2.8054	1.0846	573	2.59	0.0099
self		-0.05973	0.009372	573	-6.37	<.0001
pov		0.2728	0.07998	573	3.41	0.0007
black		0.2225	0.1243	573	1.79	0.0741
hispanic		-0.2368	0.1369	573	-1.73	0.0842
childage		0.06082	0.09018	573	0.67	0.5003
married		-0.03741	0.1254	573	-0.30	0.7656
gender		-0.4967	0.1054	573	-4.71	<.0001
momage		-0.01508	0.02505	573	-0.60	0.5473
momwork		0.2664	0.1136	573	2.35	0.0193
time	1	-0.2155	0.06295	573	-3.42	0.0007
time	2	-0.1686	0.05935	573	-2.84	0.0047
time	3	0	.	.	.	.