Longitudinal Data Analysis
Using Stata
Paul D. Allison, Ph.D.

Upcoming Seminar:
June 15-16, 2018, Boston, Massachusetts
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 estimating equations
   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 resolve causal ordering: Does $y$ cause $x$ or does $x$ cause $y$?
   Accomplished by simultaneous estimation of models with lagged predictors.
   Methods for doing this are still relatively undeveloped and underutilized.

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 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 14, 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.

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:

\[
y_{it} = \mu_i + \beta x_{it} + \gamma z_i + \epsilon_{it}, \quad i=1,\ldots, n; \ t=1,\ldots, T
\]

where \( \epsilon \) is a random error term with mean 0 and constant variance, assumed to be uncorrelated with \( x \) and \( z \). \( \beta \) and \( \gamma \) 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.
- Estimated standard errors will be too low because \( \text{corr}(\epsilon_{it}, \epsilon_{it'}) \neq 0 \)

Example:

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.

We can convert the data into a set of 1743 records, one for each child in each year using the reshape command:

```
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
Problems:

Although the coefficients are unbiased, they are not “efficient.” An estimator is said to be efficient if it has minimal sampling variability. The true standard errors are optimally small.

More important, estimated standard errors and $p$-values are probably too low.

**Solution 1: Robust standard errors**

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:
where $X$ is a matrix of dimension $Tn \times K$ (the number of coefficients) and $s^2$ is the residual variance. 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

$$
\hat{V} = (X'X)^{-1}\left(\sum_i X'_i\hat{u}_i\hat{u}'_iX_i\right)(X'X)^{-1}
$$

where $X_i$ is a $T \times K$ matrix of covariate values for individual $i$ and

$$
\hat{u}_i = y_i - X_i\hat{\beta}
$$

is a $T \times 1$ vector of residuals for individual $i$. The robust standard errors are the square roots of the main diagonal elements of $\hat{V}$.

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

```stata
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)
```
| anti | Robust Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------|-------------|-----------|-------|--------|----------------------|
| self | -0.0741425  | 0.0133707 | -5.55 | 0.000  | -0.1004034  -0.0478816 |
| pov  | 0.4354025   | 0.1093637 | 3.98  | 0.000  | 0.2206054   0.6501995 |
| black| 0.1678622   | 0.1309221 | 1.28  | 0.200  | -0.0892769  0.4250014 |
| hispanic | -0.2483772 | 0.1341785 | -1.85 | 0.065  | -0.5119122  0.0151578 |
| childage | 0.087056   | 0.0939055 | 0.93  | 0.354  | -0.0973804  0.2714923 |
| married | -0.088875  | 0.1336839 | -0.66 | 0.506  | -0.3514509  0.173676 |
| momage | -0.0166933 | 0.0241047 | -0.69 | 0.489  | -0.0640364  0.0306498 |
| gender | -0.4950259 | 0.1057334 | -4.68 | 0.000  | -0.7026929  -0.2873589 |
| momwork | 0.2120961  | 0.1189761 | 1.78  | 0.075  | -0.0215803  0.4457725 |
| year | 92 | 0.0521538  | 0.0540096 | 0.97  | 0.335  | 0.0539244   0.158232 |
| 94  | 0.2255775  | 0.0641766 | 3.51  | 0.000  | 0.0995306   0.3516245 |
| _cons | 2.675312   | 1.138426  | 2.35  | 0.019  | 0.4393717   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:
- It’s possible for robust standard errors to 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.
- Robust SEs are also robust to heteroskedasticity.
- For small samples, robust standard errors may be inaccurate and have low power. To get reasonably accurate results, you need at least 20 clusters if they are approximately balanced, 50 if they are unbalanced.

Solution 2: Generalized Estimating Equations (GEE, population averaged models)

For linear models, this is equivalent to feasible generalized least squares (GLS).
The attraction of this method is that 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 estimates.

Conventional least squares estimates are given by the matrix formula

\[(X'X)^{-1}X'y\]

GLS estimates are obtained by

\[(X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y\]

where \(\hat{\Omega}\) 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

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

In Stata, the method can be implemented with the \texttt{xtgee} command. It’s convenient to first declare the data set to be a time-series cross-section data set using the \texttt{xtset} command.

\texttt{xtset id year}

<table>
<thead>
<tr>
<th>panel variable:</th>
<th>id (strongly balanced)</th>
</tr>
</thead>
<tbody>
<tr>
<td>time variable:</td>
<td>year, 90 to 94, but with gaps</td>
</tr>
<tr>
<td>delta:</td>
<td>1 unit</td>
</tr>
</tbody>
</table>
xtgee anti self pov black hispanic childage married
gender momage momwork i.year

GEE population-averaged model

<table>
<thead>
<tr>
<th>Group variable:</th>
<th>id</th>
<th>Number of obs</th>
<th>1743</th>
</tr>
</thead>
<tbody>
<tr>
<td>Link:</td>
<td>identity</td>
<td>Obs per group: min</td>
<td>3</td>
</tr>
<tr>
<td>Family:</td>
<td>Gaussian</td>
<td>avg</td>
<td>3.0</td>
</tr>
<tr>
<td>Correlation:</td>
<td>exchangeable</td>
<td>max</td>
<td>3</td>
</tr>
<tr>
<td>Scale parameter:</td>
<td>2.275542</td>
<td>Wald chi2(11)</td>
<td>105.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prob &gt; chi2</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

| anti | Coef. | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|------|-------|-----------|-----|------|----------------------|
| self | -0.0620764 | 0.0094874 | -6.54 | 0.000 | -.0806715 to -.0434814 |
| pov  | 0.2471376  | 0.080136  | 3.08 | 0.002 | 0.090074 to 0.4042013  |
| black| 0.2267537  | 0.1249995 | 1.81 | 0.070 | -0.018241 to 0.4717483 |
| hispanic | -0.2182088 | 0.137456 | -1.59 | 0.112 | -0.4876177 to 0.0512001 |
| childage | 0.0884559 | 0.0905831 | 0.98 | 0.329 | -0.0890836 to 0.2659955 |
| married | -0.0495647 | 0.1257172 | -0.39 | 0.693 | -0.295966 to 0.1968365 |
| gender | -0.4834488 | 0.1059245 | -4.56 | 0.000 | -0.6910571 to -0.2758405 |
| momage | -0.0219197 | 0.0251467 | -0.87 | 0.383 | -0.0712064 to 0.0273669 |
| momwork | 0.2611318 | 0.1140581 | 2.29 | 0.022 | 0.037582 to 0.4846815 |
| year | 0.0473396 | 0.0585299 | 0.81 | 0.419 | -0.0673769 to 0.162056 |
| 92   | 0.2163811 | 0.0587023 | 3.69 | 0.000 | 0.1013267 to 0.3314355 |
| _cons | 2.531431 | 1.089759 | 2.32 | 0.020 | 0.3955422 to 4.667321 |

By default, the standard errors are “model based”. Although corrected for dependence, they are sensitive to the particular correlation structure that is specified.

The default correlation structure is “exchangeable”, which means that the correlations between the dependent variables at different points in time are all the same. To see the estimated correlations, use the command:

estat wcorr
Estimated within-id correlation matrix $R$:

$$
\begin{array}{c|ccc}
\hline
& c1 & c2 & c3 \\
\hline
r1 & 1 & & \\
r2 & .5636779 & 1 & \\
r3 & .5636779 & .5636779 & 1 \\
\hline
\end{array}
$$

To get robust standard errors (that aren’t sensitive to the correlation structure), simply add the robust option to the `xtgee` command:

```
xtgee anti self pov black hispanic childage married momage gender momwork i.year, vce(robust)
```

```plaintext
GEE population-averaged model
Number of obs = 1743
Group variable: id Number of groups = 581
Link: identity Obs per group: min = 3
Family: Gaussian avg = 3.0
Correlation: exchangeable max = 3
Wald chi2(11) = 90.65
Scale parameter: 2.275542 Prob > chi2 = 0.0000

(Std. Err. adjusted for clustering on id)
```

| anti | Coef. | Std. Err. | z    | P>|z|  | [95% Conf. Interval] |
|------|-------|-----------|------|------|-----------------------|
| self | -0.0620764 | 0.0101609 | -6.11 | 0.000 | -0.0819915 -0.0421614 |
| pov  | 0.2471376   | 0.0835503 | 2.96  | 0.003 | 0.0833821 0.4108932  |
| black| 0.2267537   | 0.130129  | 1.74  | 0.081 | -0.0282945 0.4818019  |
| hispanic | -0.2182088 | 0.1337172 | -1.63 | 0.103 | -0.4802896 0.043872  |
| childage | 0.0884559   | 0.0939841 | 0.94  | 0.347 | -0.0957496 0.2726615  |
| married | -0.0495647  | 0.1341853 | -0.37 | 0.712 | -0.3125631 0.2134336  |
| momage | -0.0219197  | 0.0239744 | -0.91 | 0.361 | -0.0689087 0.0250693  |
| gender | -0.4834488  | 0.1058324 | -4.57 | 0.000 | -0.6908764 -0.2760212 |
| momwork | 0.2611318    | 0.1163266 | 2.24  | 0.025 | 0.0331359 0.4891276  |
| year | 92 | 0.0473396   | 0.0535429 | 0.88  | 0.377 | -0.0576025 0.1522817  |
| 94 | 0.2163811   | 0.0634953 | 3.41  | 0.001 | 0.0919327 0.3408295  |
| _cons | 2.531431    | 1.128098  | 2.24  | 0.025 | 0.3203999 4.742463   |
```
With only three time points, you’re probably better off specifying an “unstructured” model that imposes no pattern on the correlation matrix:

```
xtgee anti self pov black hispanic childdage married
    momage gender momwork i.year, vce(r) corr(uns)
```

```
GEE population-averaged model                  Number of obs      =      1743
Group and time vars:                        id year      Number of groups   =       581
Link:                                        identity      Obs per group: min =         3
Family:                                       Gaussian                     avg =       3.0
Correlation:                  unstructured                     max =         3
                Wald chi2(11) = 94.51
Scale parameter:                  2.273983      Prob > chi2        =    0.0000
(Std. Err. adjusted for clustering on id)

<table>
<thead>
<tr>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>anti</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>self</td>
</tr>
<tr>
<td>pov</td>
</tr>
<tr>
<td>black</td>
</tr>
<tr>
<td>hispanic</td>
</tr>
<tr>
<td>childdage</td>
</tr>
<tr>
<td>married</td>
</tr>
<tr>
<td>momage</td>
</tr>
<tr>
<td>gender</td>
</tr>
<tr>
<td>momwork</td>
</tr>
<tr>
<td>year</td>
</tr>
<tr>
<td>92</td>
</tr>
<tr>
<td>94</td>
</tr>
<tr>
<td>_cons</td>
</tr>
</tbody>
</table>
```

```
estat wcorr
```
Estimated within-id correlation matrix $R$:

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>0.5512489</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>r3</td>
<td>0.5193459</td>
<td>0.6186195</td>
<td>1</td>
</tr>
</tbody>
</table>

With many time points the number of unique correlations will get large: $T(T-1)/2$. And unless the sample is also large, estimates of all these parameters may be unreliable.

In that case, consider restricted models:

<table>
<thead>
<tr>
<th>TYPE</th>
<th>Description</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR#</td>
<td>Autoregressive of order #</td>
<td>$\varepsilon_{it} = \sum_{j=1}^{#} \theta_j \varepsilon_{i,t-j} + v_{it}$</td>
</tr>
<tr>
<td>STA#</td>
<td>Stationary of order #</td>
<td>$\rho_{ts} = \rho_{s-t}$ when $</td>
</tr>
<tr>
<td>NON#</td>
<td>Non-stationary of order #</td>
<td>$\rho_{ts} = \rho_{ts}$ when $</td>
</tr>
</tbody>
</table>

Results will often be robust to choice of correlation structure, but sometimes it can make a big difference. An autoregressive structure of order 1 is usually too restrictive: the correlation goes down too rapidly with the time distance.

GEE can handle missing data on the response variable (or unbalanced panels) under the assumption that the data are missing completely at random, or that missingness depends only on the predictors. It does not allow missingness on $y$ at one time to depend on observed values of $y$ at other times.