Longitudinal Data Analysis for Social Science Researchers

Introduction to Panel Models

www.longitudinal.stir.ac.uk

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL | UNIVERSITY OF STIRLING | University of St Andrews | University of Strathclyde Glasgow

SIMPLE TABLE – PANEL DATA

Women in their 20s in 1991 – Ten Years Later

Marital status WAVE J * Marital status WAVE A Crosstabulation

Count

		Marital status WAVE A						
		Married	Living as couple	Widowed	Divorced	Separated	Never married	Total
Marital status WAVE J	Married	324	74	0	4	9	102	513
	Living as couple	16	33	1	5	6	44	105
	Widowed	4	1	1	0	1	0	7
	Divorced	36	5	0	4	9	3	57
	Separated	12	1	0	0	2	5	20
	Never married	1	18	0	0	0	85	104
Total		393	132	2	13	27	239	806

BEWARE

BEWARE

- Traditionally used in social mobility work
- Can be made more exotic for example by incorporating techniques from loglinear modelling (there is a large body of methodological literature in this area)

Change Score

- Less likely to use this approach in mainstream social science
- Understanding this will help you understand the foundation of more complex panel models (especially this afternoon)

Vector of explanatory variables and estimates

$$Y_{i1} = \underline{\beta}' \underline{x}_{i1} + \varepsilon_{i1}$$

Outcome 1 for individual i

Independent identifiably distributed error

EQUATION FOR TIME POINT 2

Vector of explanatory variables and estimates

$$Y_{i1} = \underline{\beta}' \underline{X}_{i1} + \varepsilon_{i1}$$

Outcome 1 for individual i

Independent identifiably distributed error

Considered together conventional regression analysis in NOT appropriate

$$Y_{i1} = \underline{\beta}' \underline{X}_{i1} + \varepsilon_{i1}$$
$$Y_{i2} = \underline{\beta}' \underline{X}_{i2} + \varepsilon_{i2}$$

Change in Score

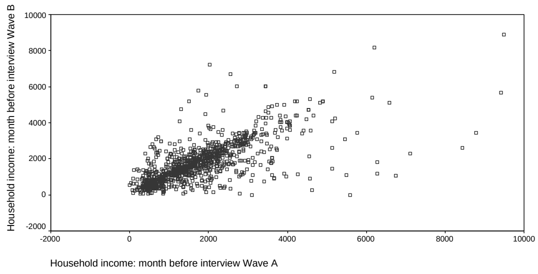
$$Y_{i2} - Y_{i1} = \underline{\beta}' (X_{i2} - X_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

Here the $\underline{\beta}'$ is simply a regression on the difference or change in scores.

Women in 20s H.H. Income Month Before Interview (Wfihhmn)

	WAVE A	WAVE B
MEAN	1793.50	1788.15
S.D.	1210.26	1171.36
MEDIAN	1566.34	1587.50
SKEWNESS	1.765	1.404
PERCENTILES		
25%	914.43	950.51
75%	2339.39	2353.85
r =.679**		

A Simple Scatter Plot



Change in Score

$$Y_{i2} - Y_{i1} = \beta'(X_{i2} - X_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

Difference or change in scores

(bfihhmn – afihhmn)

Here the β' is simply a regression on the difference or change in scores.

Models for Multiple Measures

Panel Models

PID	WAVE	SEX	AGE	Y
001	1	1	20	1
001	2	1	21	1
001	3	1	22	1
001	4	1	23	0
001	5	1	24	0
001	6	1	25	1

Models for Multiple Repeated Measures

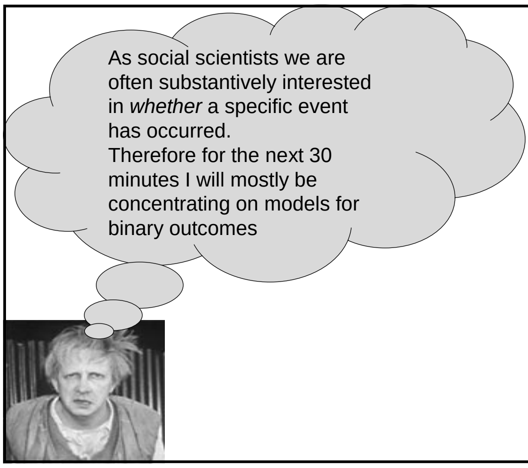
"Increasingly, I tend to think of these approaches as falling under the general umbrella of generalised linear modelling (glm). This allows me to think about longitudinal data analysis simply as an extension of more familiar statistical models from the regression family. It also helps to facilitate the interpretation of results."

Some Common Panel Models

(STATA calls these Cross-sectional time series!)

Binary Y	Logit	xtlogit
	(Probit	xtprobit)
Count Y	Poisson	xtpoisson
	(Neg bino	xtnbreg)
Continuous Y	Regression	xtreg

Models for Multiple Measures



Recurrent events are merely outcomes that can take place on a number of occasions. A simple example is unemployment measured month by month. In any given month an individual can either be employed or unemployed. If we had data for a calendar year we would have twelve discrete outcome measures (i.e. one for each month).

Consider a binary outcome or two-state event

0 = Event has not occurred
1 = Event has occurred

In the cross-sectional situation we are used to modelling this with logistic regression.

UNEMPLOYMENT AND RETURNING TO WORK STUDY –

A study for six months

0 = Unemployed; 1 = Working

Months

	1	2	3	4	5	6
obs	0	0	0	0	0	0

Constantly unemployed

Months

	1	2	3	4	5	6
obs	1	1	1	1	1	1

Constantly employed

Months

	1	2	3	4	5	6
obs	1	0	0	0	0	0

Employed in month 1
then unemployed

Months

	1	2	3	4	5	6
obs	0	0	0	0	0	1

Unemployed but gets a job in month six

Here we have a binary outcome – so could we simply use logistic regression to model it?

Yes and No – We need to think about this issue.

POOLED CROSS-SECTIONAL LOGIT MODEL

$$L^{B}{}_{it}(\underline{\beta}) = \frac{[\exp(\underline{\beta}'\underline{x}_{it})]^{y_{it}}}{1+\exp(\underline{\beta}'\underline{x}_{it})}$$

$\underline{x}_{it}$ is a vector of explanatory variables and $\underline{\beta}$ is a vector of parameter estimates

We could fit a pooled cross-sectional model to our recurrent events data.

This approach can be regarded as a naïve solution to our data analysis problem.

We need to consider a number of issues….

Months

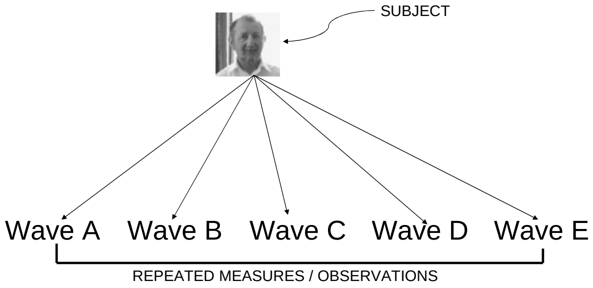
	Y_1	Y_2
obs	0	0

Pickle's tip - In repeated measures analysis we would require something like a 'paired' t test rather than an 'independent' t test because we can assume that Y_1 and Y_2 are related.

Repeated measures data violate an important assumption of conventional regression models.

The responses of an individual at different points in time will not be independent of each other.

The observations are "clusters" in the individual



Repeated measures data violate an important assumption of conventional regression models.

The responses of an individual at different points in time will not be independent of each other.

This problem has been overcome by the inclusion of an additional, individual-specific error term.

POOLED CROSS-SECTIONAL LOGIT MODEL

$$L^{B}{}_{it}(\underline{\beta}) = \frac{[\exp(\underline{\beta}'\underline{x}_{it})]^{y_{it}}}{1 + \exp(\underline{\beta}'\underline{x}_{it})}$$

PANEL LOGIT MODEL (RANDOM EFFECTS MODEL)
Simplified notation!!!

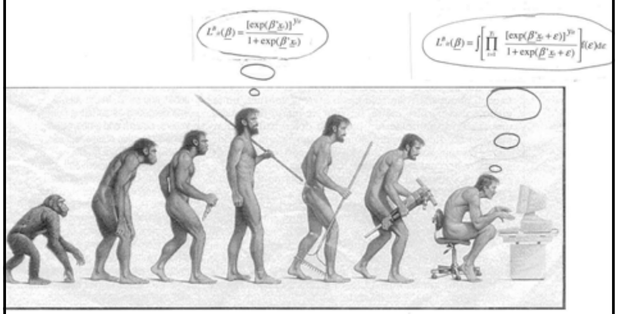
$$L^{B}{}_{it}(\underline{\beta}) = \int \left[\prod_{t=1}^{T_i} \frac{[\exp(\underline{\beta}'\underline{x}_{it} + \varepsilon)]^{y_{it}}}{1 + \exp(\underline{\beta}'\underline{x}_{it} + \varepsilon)} \right] \mathrm{f}(\varepsilon)\mathrm{d}\varepsilon$$

For a sequence of outcomes for the i^{th} case, the basic random effects model has the integrated (or marginal likelihood) given by the equation.

$$L^{B}{}_{it}(\underline{\beta}) = \int \left[\prod_{t=1}^{T_i} \frac{[\exp(\underline{\beta}'\underline{x}_{it} + \varepsilon)]^{y_{it}}}{1 + \exp(\underline{\beta}'\underline{x}_{it} + \varepsilon)} \right] \mathrm{f}(\varepsilon)\mathrm{d}\varepsilon$$

The random effects model extends the pooled cross-sectional model to include a case-specific random error term this helps to account for residual heterogeneity.

Davies and Pickles (1985) have demonstrated that the failure to explicitly model the effects of residual heterogeneity may cause severe bias in parameter estimates. Using longitudinal data the effects of omitted explanatory variables can be overtly accounted for within the statistical model. This greatly improves the accuracy of the estimated effects of the explanatory variables.



An simple example –
Davies, Elias & Penn (1992)

The relationship between a husband's unemployment and his wife's participation in the labour force

Four waves of BHPS data

Married Couples in their 20s (n=515; T=4; obs=2060)

Summary information…

56% of women working (in paid employment)

59% of women with employed husbands work
23% of women with unemployed husbands work

65% of women without a child under 5 work
48% of women with a child under 5 work

POOLED (cross-sectional) MODELS

MODEL	**Deviance (Log L)**	**Change d.f.**	**No. obs**
Null Model	2830 (-1415)	-	2060
+ husband unemployed	2732 (-1366)	1	2060
+husband u +child und 5	2692 (-1346)	1	2060
husband u * child und 5	2692 (-1346)	1	2060

First glimpse at STATA

- Models for panel data
- STATA – unhelpfully calls this ‘cross-sectional time-series’
- *xt* commands suite

STATA CODE

Cross-Sectional Model

logit y mune und5

Cross-Sectional Model

logit y mune und5, cluster (pid)

POOLED MODELS

	Cross-sectional (pooled)		**Cross-sectional (cluster)**	
	Beta	S.E.	Beta	S.E.
Husband unemployed	-1.49	0.18	-1.49	0.23
Child under 5	-0.59	0.09	-0.59	0.13
Constant	0.69	0.07	0.69	0.11

xtdes, i(pid) t(year)

```
xtdes, i(pid) t(year)

    pid:  10047093, 10092986, ..., 19116969                n =      515
    year:  91, 92, ..., 94                                 T =        4
          Delta(year) = 1; (94-91)+1 = 4
          (pid*year uniquely identifies each observation)

Distribution of T_i:   min     5%     25%      50%       75%     95%     max
                         4      4      4        4         4       4       4

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      515    100.00  100.00 |  1111
 ---------------------------+---------
      515    100.00         |  XXXX
```

xtdes, i(pid) t(year)

```
xtdes, i(pid) t(year)

    pid:  10047093, 10092986, ..., 19116969                n =       515
    year:  91, 92, ..., 94                                 T =         4
```

xtdes, i(pid) t(year)

```
Delta(year) = 1; (94-91)+1 = 4
         (pid*year uniquely identifies each observation)
```

xtdes, i(pid) t(year)

```
Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                          4      4       4         4         4       4       4
```

xtdes, i(pid) t(year)

Note: this is a balanced panel

```
  Freq.  Percent    Cum. |  Pattern
---------------------------+---------
    515    100.00  100.00 |  1111
---------------------------+---------
    515    100.00             |  XXXX
```

xtdes, i(pid) t(year)

```
xtdes, i(pid) t(year)

     pid:  10047093, 10092986, ..., 19116969                   n =       515
    year:  91, 92, ..., 94                                     T =         4
           Delta(year) = 1; (94-91)+1 = 4
           (pid*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         4       4       4         4         4       4       4

  Freq.  Percent    Cum. |  Pattern
---------------------------+---------
    515    100.00  100.00 |  1111
---------------------------+---------
    515    100.00             |  XXXX
```

POOLED & PANEL MODELS

MODEL	**Deviance (Log L)**	**Change d.f.**	**No. obs**
Pooled Model	2692 (-1346)	-	2060
Panel Model	2186 (-1093)	1	2060 (n=515)

The panel model is clearly an improvement on the pooled cross-sectional analysis. We can suspect non-independence of observations.

FURTHER - EXPLORATION

PANEL MODELS

MODEL	**Deviance (Log L)**	**Change d.f.**	**No. obs (n)**
Null Model	2218 (-1109)	-	2060 (515)
+ husband unemployed	2196 (-1098)	1	2060 (515)
+husband u +child und 5	2186 (-1093)	1	2060 (515)
husband u * child und 5	2186 (-1093)	1	2060 (515)

COMPARISON OF MODELS

	Cross-sectional (pooled)			**Random Effects**	
	Beta	S.E.	Rob S.E.	Beta	S.E.
Husband unemployed	-1.49	0.18	0.23	-.83	.18
Child under 5	-0.59	0.09	0.13	-.34	.10
Constant	0.69	0.07	0.11	.53	.10

STATA OUTPUT

```
Random-effects logistic regression           Number of obs      =      2060
Group variable (i): pid                      Number of groups   =       515
Random effects u_i ~ Gaussian                Obs per group: min =         4
                                              avg =       4.0
                                              max =         4
                                              Wald chi2(2)       =    31.73
Log likelihood  = -1093.3383                 Prob > chi2        =   0.0000
------------------------------------------------------------------------------
         y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Imune_1 |  -1.351039   .3029752    -4.46   0.000    -1.944859   -.7572184
 _Iund5_1 |  -.5448233   .1712375    -3.18   0.001    -.8804426    -.209204
     _cons |   .8551312   .1557051     5.49   0.000     .5499549    1.160307
-------------+----------------------------------------------------------------
  /lnsig2u |   1.659831   .0974218                      1.468888    1.850774
-------------+----------------------------------------------------------------
   sigma_u |   2.293125   .1117002                      2.084322    2.522845
       rho |   .6151431   .0230638                      .5690656    .6592439
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   504.79 Prob >= chibar2 = 0.000
```


```
Random-effects logistic regression           Number of obs      =      2060
Group variable (i): pid                      Number of groups   =       515
Random effects u_i ~ Gaussian                Obs per group: min =         4
                                              avg =       4.0
                                              max =         4
                                              Wald chi2(2)       =    31.73
Log likelihood  = -1093.3383                 Prob > chi2        =   0.0000
------------------------------------------------------------------------------
         y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Imune_1 |  -1.351039   .3029752    -4.46   0.000    -1.944859   -.7572184
 _Iund5_1 |  -.5448233   .1712375    -3.18   0.001    -.8804426    -.209204
     _cons |   .8551312   .1557051     5.49   0.000     .5499549    1.160307
-------------+----------------------------------------------------------------
  /lnsig2u |   1.659831   .0974218                      1.468888    1.850774
-------------+----------------------------------------------------------------
   sigma_u |   2.293125   .1117002                      2.084322    2.522845
       rho |   .6151431   .0230638                      .5690656    .6592439
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   504.79 Prob >= chibar2 = 0.000
```

```
Random-effects logistic regression          Number of obs      =      2060
Group variable (i): pid                     Number of groups   =       515
Random effects u_i ~ Gaussian               Obs per group: min =         4
                                                           avg =       4.0
                                                           max =         4
                                            Wald chi2(2)       =     31.73
Log likelihood  = -1093.3383                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imune_1 |  -1.351039   .3029752    -4.46   0.000    -1.944859   -.7572184
   _Iund5_1 |  -.5448233   .1712375    -3.18   0.001    -.8804426    -.209204
      _cons |   .8551312   .1557051     5.49   0.000     .5499549    1.160307
-------------+----------------------------------------------------------------
    /lnsig2u |   1.659831   .0974218                      1.468888    1.850774
-------------+----------------------------------------------------------------
     sigma_u |   2.293125   .1117002                      2.084322    2.522845
         rho |   .6151431   .0230638                      .5690656    .6592439
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   504.79 Prob >= chibar2 = 0.000
```


```
Random-effects logistic regression          Number of obs      =      2060
Group variable (i): pid                     Number of groups   =       515
Random effects u_i ~ Gaussian               Obs per group: min =         4
                                                           avg =       4.0
                                                           max =         4
                                            Wald chi2(2)       =     31.73
Log likelihood  = -1093.3383                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Imune_1 |  -1.351039   .3029752    -4.46   0.000    -1.944859   -.7572184
   _Iund5_1 |  -.5448233   .1712375    -3.18   0.001    -.8804426    -.209204
      _cons |   .8551312   .1557051     5.49   0.000     .5499549    1.160307
-------------+----------------------------------------------------------------
    /lnsig2u |   1.659831   .0974218                      1.468888    1.850774
-------------+----------------------------------------------------------------
     sigma_u |   2.293125   .1117002                      2.084322    2.522845
         rho |   .6151431   .0230638                      .5690656    .6592439
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   504.79 Prob >= chibar2 = 0.000
```


```
-------------+----------------------------------------------------------------
    /lnsig2u |   1.659831   .0974218                      1.468888    1.850774
-------------+----------------------------------------------------------------
     sigma_u |   2.293125   .1117002                      2.084322    2.522845
         rho |   .6151431   .0230638                      .5690656    .6592439
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   504.79 Prob >= chibar2 = 0.000
```

sigma_u can be interpreted like a parameter estimate with a standard error

Remember it is the root of anti log - sig2u

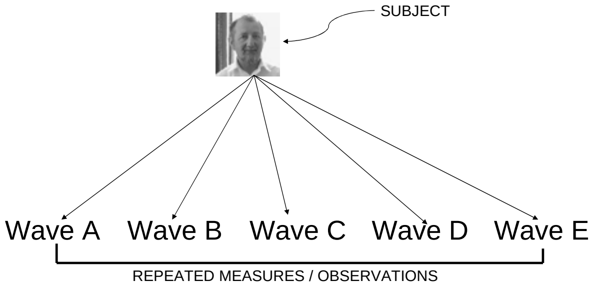
$$\sqrt{(\exp\sigma^2)}$$

```
-------------+----------------------------------------------------------------
    /lnsig2u |   1.659831   .0974218                      1.468888    1.850774
-------------+----------------------------------------------------------------
     sigma_u |   2.293125   .1117002                      2.084322    2.522845
         rho |   .6151431   .0230638                      .5690656    .6592439
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =   504.79 Prob >= chibar2 = 0.000
```

rho = sigma_u / (sigma_u + sigma_e)
rho can be appreciated as the proportion of the total variance contributed by the panel-level (i.e. subject level) variance component

When rho is zero the panel-level variance component is unimportant. A likelihood ratio test is provided at the bottom of the output

You can think of rho as being the (analogous) equivalent of the intra-cluster correlation (icc) in a multilevel model



SOME CONCLUSIONS

- Panel models are attractive
- Extend cross-sectional (glm) models
- They overcome the non-independence problem
- Provide increased control for residual heterogeneity
- Can be extended to provide increased control for state dependence

Some Restrictions

- Specialist software (e.g. STATA)
- {Powerful computers required}
- Results can be complicated to interpret
- Real data often behaves badly (e.g. unbalanced panel)
- Communication of results can be more tricky