

# Longitudinal/Panel Data Analysis

Raymond Duch

University of Oxford  
Nuffield College  
[raymond.duch@nuffield.ox.ac.uk](mailto:raymond.duch@nuffield.ox.ac.uk)  
[raymond Duch.com/trinity10/paneldata](http://raymond Duch.com/trinity10/paneldata)

April 27, 2010

- 1 Gellman, Andrew and Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press
- 2 Stata 11.0 Manual Longitudinal/Panel Data
- 3 Rabe-Hesketh, Sophia and Anders Skrondal. 2005. Multilevel and Longitudinal Modeling Using Stata. Stata Press

# What is longitudinal panel data?

- 1 Marriage of regression and time-series analysis
- 2 A broad cross-section of subjects observed over time
- 3 Individuals surveyed repeatedly over time (American National Election Study; U.S. Panel Study of Income Dynamics)
- 4 Statistics compiled over time for a particular geo-political entity (Divorce Rates and welfare rates collected annually from U.S. States)
- 5 Statistics compiled on hospital patients over time

(Repeated) cross-sectional regression analysis generates the following model

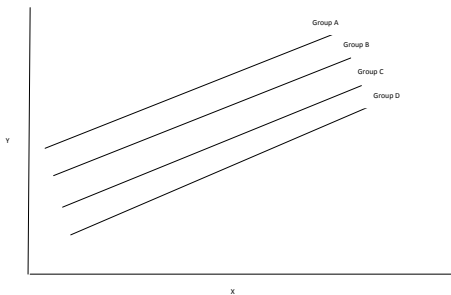
$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it} \quad (1)$$

$$y_{it} = \alpha + \mathbf{x}'_{it} \mathbf{B} + \epsilon_{it} \quad (2)$$

- 1 Heterogeneity or uniqueness of subjects captured in  $\epsilon_{it}$
- 2 The cross-sectional units (individuals, firms, cities) are represented by  $i$
- 3 Repeated time units are represented by  $t$

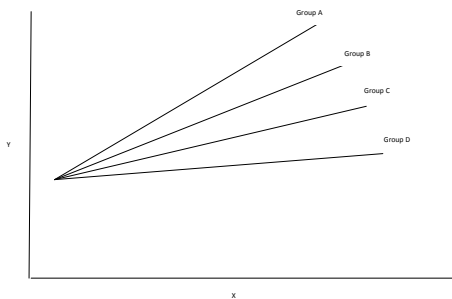
# Varying Intercept Model

$$y_{it} = \alpha_j + \beta x_{it} + \epsilon_{it} \quad (3)$$



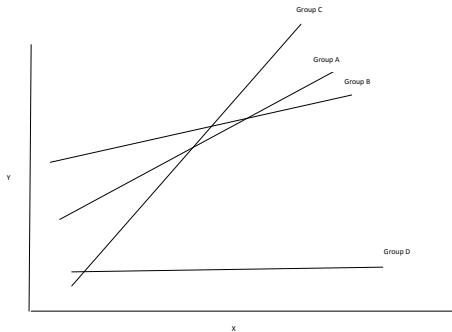
# Varying Slope Model

$$y_{it} = \alpha + \beta_j x_{it} + \epsilon_{it} \quad (4)$$



# Varying Intercepts and Slopes Model

$$y_{it} = \alpha_j + \beta_j x_{it} + \epsilon_{it} \quad (5)$$

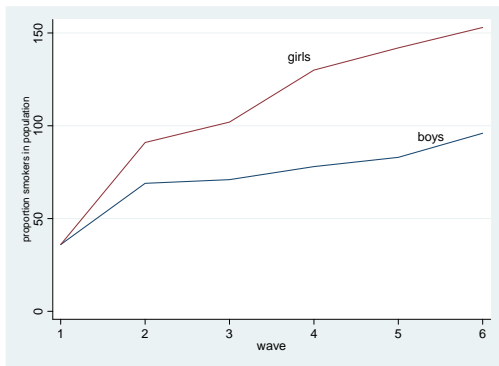


- 1 data is available at  
<http://www.stat.columbia.edu/gelman/arm/>
- 2 variables: newid (identifies each unique respondent) sex (1=female) parsmk (1=parents smoke) wave (identifies each of 6 waves) smkreg (is respondent regular smoker)

. list

```
+-----+
| newid  sex_1_f_  parsmk  wave  smkreg |
+-----+
1. |      1          1          0      1      0 |
2. |      1          1          0      2      0 |
3. |      1          1          0      4      0 |
4. |      1          1          0      5      0 |
5. |      1          1          0      6      0 |
+-----+
6. |      2          0          0      1      0 |
7. |      2          0          0      2      0 |
8. |      2          0          0      3      0 |
9. |      2          0          0      4      0 |
10. |     2          0          0      5      0 |
+-----+
11. |     2          0          0      6      0 |
12. |     3          1          0      1      0 |
13. |     3          1          0      2      0 |
14. |     3          1          0      3      0 |
15. |     3          1          0      4      0 |
+-----+
16. |     3          1          0      5      0 |
17. |     3          1          0      6      0 |
18. |     4          1          0      1      0 |
19. |     4          1          0      2      0 |
20. |     4          1          0      3      0 |
+-----+
21. |     4          1          0      4      0 |
22. |     4          1          0      5      0 |
23. |     4          1          0      6      0 |
24. |     5          0          0      1      0 |
25. |     5          0          0      2      0 |
+-----+
26. |     5          0          0      3      0 |
27. |     5          0          0      4      0 |
28. |     5          0          0      5      0 |
```

# Smoking by Sex over Panel Waves



$$Pr(y_{jt} = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 \text{psmoke}_{jt} + \beta_2 \text{female}_{jt} + \quad (6)$$

$$\beta_3 t + \beta_4 \text{female}_{jt} * t + \alpha_j + \epsilon_{jt}), t = 1, \dots, T_j, j = 1, \dots, n. \quad (7)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \quad (8)$$

# Estimation with Gllamm in Stata

```
. use "e:\Oxford08\Department08\Trinity_Panel\Data\gelman\smoke_pub.dta", clear
.
. tsset newid wave
    panel variable:  newid (unbalanced)
    time variable:   wave, 1 to 6, but with gaps
                   delta: 1 unit
. gllamm smkreg parsmk wave, i(newid) link(logit) family(binom)
```

# Estimation with Gllamm in Stata

```
gllamm model
```

```
log likelihood = -2074.7563
```

```
-----+-----  
smkreg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]  
-----+-----  
  parsmk |   1.270422   .1998237     6.36   0.000     .8787746   1.662069  
    wave |   .4195264   .0365132    11.49   0.000     .3479619   .4910909  
    _cons |  -7.24026   .2742149   -26.40   0.000    -7.777711  -6.702808  
-----+-----
```

```
Variiances and covariances of random effects
```

```
***level 2 (newid)
```

```
var(1): 13.679018 (.88531601)
```

# Estimation with Gllamm in Stata: Incorporating Time Trend

```
.  
. gen male_time=wave*(1-sex_1_f)  
. gen female_time=wave*sex_1_f  
. gen sex_time=wave*sex_1_f  
. gllamm smkreg parsmk wave sex_time, i(newid) link(logit) family(binom)
```

# Estimation with Gllamm in Stata: Incorporating Time Trend

```
.  
. gen male_time=wave*(1-sex_1_f)  
. gen female_time=wave*sex_1_f  
. gen sex_time=wave*sex_1_f  
. gllamm smkreg parsmk wave sex_time, i(newid) link(logit) family(binom)
```

# Estimation with Gllamm in Stata: Incorporating Time Trend

number of level 1 units = 8730

number of level 2 units = 1760

Condition Number = 17.565231

gllamm model

log likelihood = -2071.4531

```
-----+-----
```

smkreg	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
parsmk	1.314832	.2278361	5.77	0.000	.8682812	1.761382
wave	.3598051	.0432529	8.32	0.000	.275031	.4445792
sex_time	.10706	.0424822	2.52	0.012	.0237965	.1903235
_cons	-7.263204	.2767673	-26.24	0.000	-7.805658	-6.72075
-----+-----						

Variances and covariances of random effects

```
-----+-----
```

\*\*\*level 2 (newid)

var(1): 13.797342 (.90193295)

```
-----+-----
```

# Estimation with Gllamm in Stata: Incorporating Time Trend

```
+-----+
| newid   constant   reffm1   inter_eb |
+-----+
1. |      1   -7.263204  -1.1592099  -8.422414 |
2. |      1   -7.263204  -1.1592099  -8.422414 |
3. |      1   -7.263204  -1.1592099  -8.422414 |
4. |      1   -7.263204  -1.1592099  -8.422414 |
5. |      1   -7.263204  -1.1592099  -8.422414 |
+-----+
```

```
. list newid constant reffm1 inter_eb in 1090/1095
```

```
+-----+
| newid   constant   reffm1   inter_eb |
+-----+
1090. |     202   -7.263204  -.76498347  -8.028188 |
1091. |     203   -7.263204   7.1519595  -.1112444 |
1092. |     203   -7.263204   7.1519595  -.1112444 |
1093. |     203   -7.263204   7.1519595  -.1112444 |
1094. |     203   -7.263204   7.1519595  -.1112444 |
+-----+
1095. |     203   -7.263204   7.1519595  -.1112444 |
+-----+
```

```
. list newid constant reffm1 inter_eb in 1160/1165
```

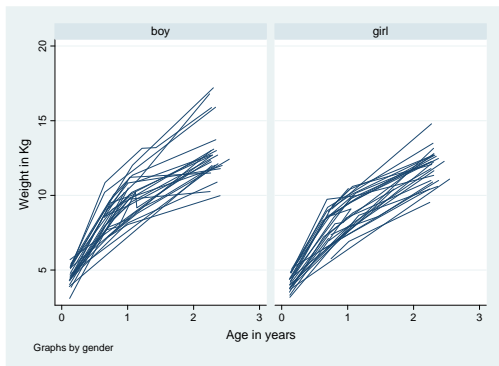
```
+-----+
| newid   constant   reffm1   inter_eb |
+-----+
1160. |     215   -7.263204   6.0917393  -1.171465 |
1161. |     215   -7.263204   6.0917393  -1.171465 |
1162. |     215   -7.263204   6.0917393  -1.171465 |
1163. |     215   -7.263204   6.0917393  -1.171465 |
1164. |     216   -7.263204  -1.4855779  -8.748782 |
+-----+
```

- 1 data is available with following command: net from <http://www.stata-press.com/data/mlmus2/>
- 2 variables: id (child identifier) weight (weight in Kg) age (age in years) gender (1 male; 2 female)

. list

	id	occ	age	weight	brthwt	gender
1.	45	1	.136893	5.171	4140	boy
2.	45	2	.657084	10.86	4140	boy
3.	45	3	1.21834	13.15	4140	boy
4.	45	4	1.42916	13.2	4140	boy
5.	45	5	2.27242	15.88	4140	boy
6.	258	1	.19165	5.3	3155	girl
7.	258	2	.687201	9.74	3155	girl
8.	258	3	1.12799	9.98	3155	girl
9.	258	4	2.30527	11.34	3155	girl
10.	287	1	.134155	4.82	3850	boy
11.	287	2	.70089	9.09	3850	boy
12.	287	3	1.16906	11.1	3850	boy
13.	287	4	2.2423	16.8	3850	boy
14.	483	1	.747433	5.76	2875	girl
15.	483	2	1.01848	6.92	2875	girl
16.	483	3	2.24504	9.53	2875	girl
17.	725	1	.120465	4.4	3280	girl
18.	725	2	2.30527	12.25	3280	girl
19.	800	1	1.12252	10.89	3900	boy
20.	800	2	2.26146	12.7	3900	boy

# Observed growth trajectories for boys and girls



$$y_{jt} = \beta_0 + \beta_1 \text{age}_{jt} + \beta_2 \text{age}_{jt}^2 + \alpha_j + \epsilon_{jt}, \quad (9)$$

$$t = 1, \dots, T_j, j = 1, \dots, n. \quad (10)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \quad (11)$$

# Estimation with xtmixed in Stata

```
. gen age2=age^2
```

```
. xtmixed weight age age2 || id:, mle
```

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log likelihood = -276.83266

Iteration 1: log likelihood = -276.83266

Computing standard errors:

Mixed-effects ML regression

Group variable: id

Number of obs = 198

Number of groups = 68

Obs per group: min = 1

avg = 2.9

max = 5

Wald chi2(2) = 2623.63

Prob > chi2 = 0.0000

Log likelihood = -276.83266

weight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	7.817918	.2896529	26.99	0.000	7.250209	8.385627
age2	-1.705599	.1085984	-15.71	0.000	-1.918448	-1.49275
_cons	3.432859	.1810702	18.96	0.000	3.077968	3.78775

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
id: Identity				
sd(_cons)	.9182256	.0973788	.7458965	1.130369
sd(Residual)	.7347063	.0452564	.6511507	.8289837

LR test vs. linear regression: chibar2(01) = 78.07 Prob >= chibar2 = 0.0000

$$y_{jt} = \beta_0 + \beta_1 age_{jt} + \beta_2 age_{jt}^2 + \beta_3 girl_{jt} + \beta_4 girl * age_{jt} \quad (12)$$

$$\alpha_j + \epsilon_{jt}, t = 1, \dots, T_j, j = 1, \dots, n. \quad (13)$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2) \quad (14)$$

# Estimation with xtmixed in Stata

```
. xtmixed weight age age2 girl age_girl || id:, mle
```

```
Iteration 1: log likelihood = -270.7967
```

```
Mixed-effects ML regression
```

```
Group variable: id
```

```
Number of obs = 198
```

```
Number of groups = 68
```

```
Obs per group: min = 1
```

```
avg = 2.9
```

```
max = 5
```

```
Log likelihood = -270.7967
```

```
Wald chi2(4) = 2705.20
```

```
Prob > chi2 = 0.0000
```

weight	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	7.932362	.2935717	27.02	0.000	7.356973	8.507752
age2	-1.70546	.1069802	-15.94	0.000	-1.915138	-1.495783
girl	-.4889737	.2752022	-1.78	0.076	-1.02836	.0504127
age_girl	-.2289743	.1377625	-1.66	0.096	-.4989839	.0410353
_cons	3.676974	.2212291	16.62	0.000	3.243373	4.110575

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
id: Identity				
sd(_cons)	.8470338	.0921964	.6843065	1.048457
sd(Residual)	.7261711	.0446575	.6437132	.8191916

```
LR test vs. linear regression: chibar2(01) = 69.16 Prob >= chibar2 = 0.0000
```

With two periods and strict exogeneity,

$$y_{it} = \beta_0 + \beta_1 D_{i2} + \beta_2 T_t + \beta_3 T_t D_{it} + \epsilon_{it} \quad (15)$$

- 1  $D_{i2}$  = dummy variable for a treatment that takes place between time 1 and time 2 for some of the individuals
- 2  $T_t$  = a time dummy variable, 0 in period 1, 1 in period 2
- 3 This is a classic regression model. If there are no regressors, using least squares,

$$\beta_3 = (\bar{y}_2 - \bar{y}_1)_{D=1} - (\bar{y}_2 - \bar{y}_1)_{D=0} \quad (16)$$

- Gellman, Andrew and Jennifer Hill. 2007. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Chapter 13 and 14
- Stata 11.0 Manual Longitudinal/Panel Data, xtmixed, xtreg, xtregar
- Rabe-Hesketh, Sophia and Anders Skrondal. 2005. Multilevel and Longitudinal Modeling Using Stata. Stata Press, Chapter 3 and 4
- Halaby, Charles. 2004. "Panel Models in Sociological Research: Theory and Practice." Annual Review of Sociology. 30: 507-44
- Wooldridge, J.M. 2002. "Econometric Analysis of Cross Section and Panel Data Cambridge, MA : MIT Press (especially chapters 13 and 14).