

Introduction to Longitudinal Data Analysis

Introduction

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN



Why Use Longitudinal Data?

- UK has an unparalleled collection
- These resources are critical for analysing social change (and social stability)
- But they need justification because they are costly in money and time

Longitudinal Social Surveys

- Cross-sectional data
 - Respondents surveyed at only one time point
- Longitudinal data
 - Repeated contacts (with the same individuals)
 - Respondents surveyed at multiple time points



Longitudinal Social Science Study Designs

Panel Study

The panel are the group and are repeatedly studied

- US (PSID)
- Germany (SOEP)
- Britain BHPS/UKHLS
- Australia (HILDA)
- Canada (SLID)
- Swiss (SHP); Korea (KLIPS); Russia (RLMS)

Longitudinal Social Science Study Designs

Cohort Study

- Repeated contacts data collection
(simply a specific form of panel design in my view)
- Principally concerned with charting the development of a particular 'group' from a certain point in time

Longitudinal Social Science Study Designs

- Cohort Study

- A birth cohort of babies born in a particular year (e.g. 1946; 1958; 1970; 2000-2)
- A youth cohort, a group of pupils who completed compulsory education in the same year (YCS; LSYPE)

Research Using Longitudinal Social Survey Datasets

- For many social research projects cross-sectional data will be sufficient
- Most social research projects can be improved by the analysis of longitudinal data
- Some research questions require longitudinal data

Questions that Require Longitudinal Data

- Flows into and out of poverty
- The effects of family migration on the woman's subsequent employment activities
- Numerous policy intervention examples
- Numerous examples relating to 'individual' development

Key Messages (so far....)

- For many social research projects cross-sectional data will be sufficient
- Most social research projects can be improved by the analysis of longitudinal data
- *Researchers are likely to make more rapid progress using existing large-scale longitudinal data resources*

Key Messages (so far....)

- Some research questions require longitudinal data
- Longitudinal data are not a panacea





‘This longitudinal study suggests that notwithstanding the dominant effect of severity of intellectual impairment, a number of factors within and outside the family may also contribute to higher attainment in reading, writing and numeracy.

In particular mainstream schooling for those with less severe disabilities appears to have benefited the children in this study’ (p.390).

Turner, S., Alborz, A. and **Gayle, V.** (2008) ‘Predictors of academic attainments of young people with Down’s syndrome’, *Journal of Intellectual Disability Research*, 52(5), pp. 380-392.

Subjective Well-Being & Happiness

- Non-economic measures of social progress
- “Improving the quality of our lives should be the ultimate target of public policies” Angel Gurría, OECD Secretary-General
- UK commitment to developing wider measures of well-being
- Tailoring government policies to the things that matter

- Moving house itself causes a boost in happiness, and brings people back to their initial levels
- Moving and set-point theory
- Long-distance migrants are at least as happy as short-distance migrants despite the higher social and psychological costs involved
- Re-theorize moving within a conceptual framework that accounts for social well-being from a life-course perspective

Nowok, B., van Ham, M., Findlay, A. and **Gayle, V.** (2013) 'Does migration make you happy? A longitudinal study of internal migration and subjective wellbeing', *Environment and Planning A*, 45(4), pp. 986-1002.

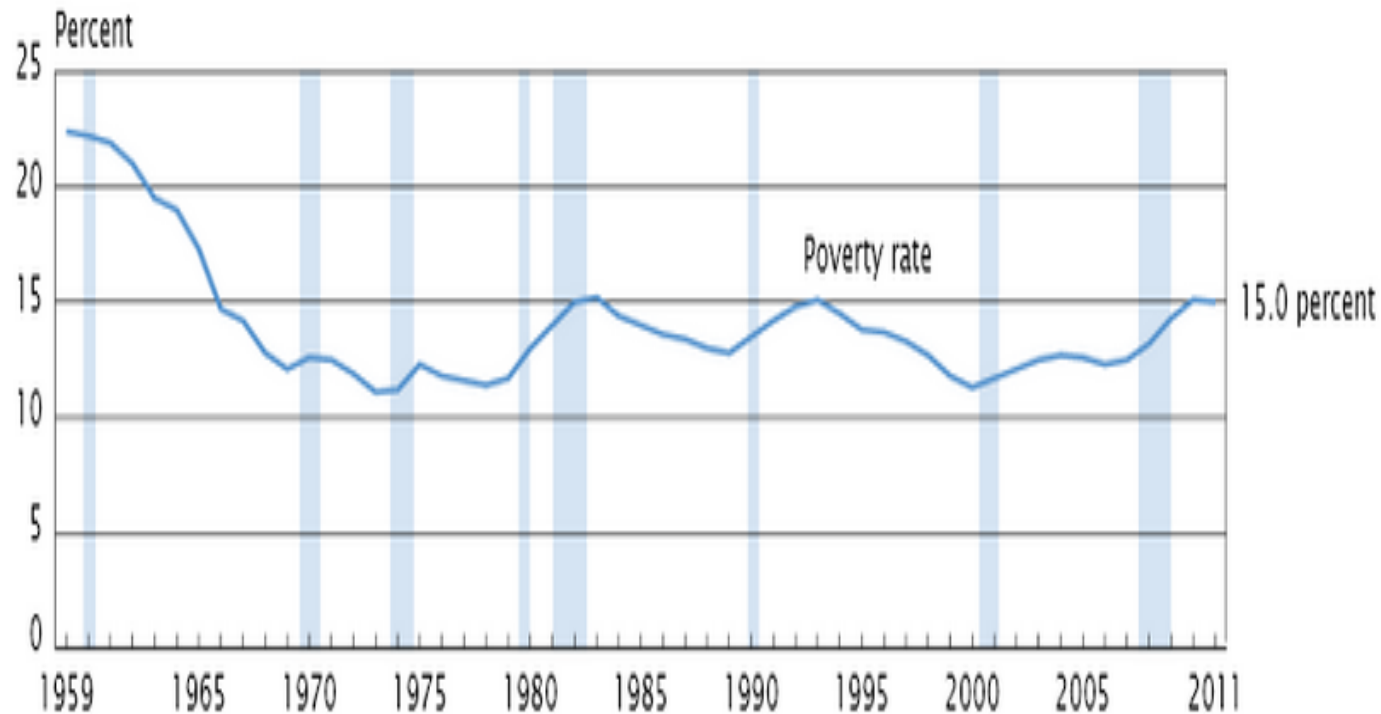
The Bigger Picture

- UKHLS is the largest living observatory of contemporary social life
- Contribution to the 'evidence base'
- Contribution to empirically informed planning
- *Influencing behaviour and informing interventions*
- *Contributing to a fair and vibrant society*

Examples...

- Cohort studies - secondary smoking effects on children (back in the news)
- Whitehall Studies - influenced successive governments' thinking on social gradients in health
- Whitehall Studies - dispelled the myth that high status jobs have higher risk of heart disease

USA Poverty Rate 1959 - 2011



Note: The data points are placed at the midpoints of the respective years. For information on recessions, see Appendix A.

Source: U.S. Census Bureau, Current Population Survey, 1960 to 2012 Annual Social and Economic Supplements.

- Poverty rates flattened out in 1990s
- BHPS showed apparent cross-sectional stability but a hidden longitudinal flux
 - Substantial turnover or churning
 - The poor were not always poor
- Not detectable without panel data!

- UK Poverty rate approximately 18%
- In a 6 year periods one-third of individuals were poor at least once
- Only 2% were poor for all six years!
- Repeated short spells of poverty were more common than one long spell

The Consequences...

- Contributed to the 'rubber band theory'
 - we are attached to an elastic tether
- Influenced the Labour government's welfare reforms in the late 1990s
 - focussing on moving people into work and making work pay
- Now influences how living standards are measured in Britain
 - Official Statistics now include household panel based information

Summary Messages

- For many social research projects cross-sectional data will be sufficient
- Most social research projects can be improved by the analysis of longitudinal data
- Some research questions require longitudinal data

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

The Research Value of Longitudinal Data

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

A vignette...

The story of Jason Jones (aged 10) and his mum



Questions that Require Longitudinal Data

- Flows into and out of poverty
- The effects of family migration on the woman's subsequent employment activities
- Numerous policy intervention examples
- Numerous examples relating to 'individual' development

Methodological Benefits of Longitudinal Social Science Data

- Micro-level social processes
- Temporal ordering of events
- Improving control for residual heterogeneity
- Improving control for state dependence

Micro-Level Social Processes

- Cross-sectional data = a snap shot
 - Good for studying the immediate
 - Several datasets can study macro / or gross changes
- Repeated contacts data allow the study of
 - The passage of time
 - Individual (or household) change/stability
 - Processes that occur at the micro-level of the individual (or family)
 - Surprises (or shocks)

Temporal Ordering of Events (Direction of Influence)

- Time moves in one direction so...
 - An event in 1990 comes before an event in 1995
 - Experiences at primary school could affect university entry
 - Teenage smoking could influence health in old age
- But not *vice versa*
 - One sociology professor has argued with me suggesting that time does not move in only one direction

Temporal Ordering of Events (Direction of Influence)

- There is unequivocal evidence from cross-sectional data that, overall, the unemployed have poorer health
- This is consistent with both
 - A. Unemployment causing ill health
 - B. Ill health causing unemployment
- These two substantive stories are quite different

Month	Level of Health (20 = Good Health)	Employ Status
1	17	Employed
2	17	Employed
3	17	Employed
4	17	Unemployed
5	17	Unemployed
6	10	Unemployed
7	16	Unemployed
8	5	Unemployed
9	4	Unemployed
10	3	Unemployed
11	2	Unemployed
12	1	Unemployed

Person A



Became unemployed this has affected
his level of health

Month	Level of Health (20 = Good Health)	Employ Status
1	17	Employed
2	1	Employed
3	1	Employed
4	1	Unemployed
5	1	Unemployed
6	1	Unemployed
7	1	Unemployed
8	1	Unemployed
9	1	Unemployed
10	1	Unemployed
11	1	Unemployed
12	1	Unemployed

Person B



Poor health led to unemployment
(because of poor job performance)

In a cross-sectional study (at month 12)

- Person A would have been unemployed for 9 months and have a health score of 1
- Person B would have been unemployed for 9 months and have a health score of 1
- This is an obvious example of how panel (i.e. repeated contacts) data can make an essential contribution to untangling social relationships

Improving Control for Omitted Explanatory Variables

- Residual Heterogeneity
 - Omitted explanatory variables
 - Unobserved heterogeneity
- The possibility of substantial variation between similar individuals due to unmeasured, and possibly immeasurable, variables is known as '*residual heterogeneity*'

Improving Control for Omitted Explanatory Variables

Because data collection instruments often fail to capture the detailed nature of social life there is, almost inevitably, considerable heterogeneity in response variables even amongst respondents that share the same characteristics across all of the explanatory variables

Improving Control for Omitted Explanatory Variables

As long as we make the assumption that (at least some of) these effects are enduring there are techniques for accounting for omitted explanatory variables if we have data at more than one time point

Improving Control for Omitted Explanatory Variables

- There are no routine methods of accounting for omitted explanatory variables in cross-sectional analysis
- It is sometimes claimed that the main advantage of longitudinal data is that it facilitates improved control for the plethora of variables that are omitted from any analysis
- Panel data won't completely sweep this problem away, but suitable models can improve control for, and estimate the effects of, residual heterogeneity

Improving Control for the Effects of Previous States (state dependence)

A frequently noted empirical regularity in the analysis of unemployment data is that those who were unemployed in the past or have worked in the past are more likely to be unemployed (or working) in the future

(Nobel Prize winner J.J. Heckman)

Improving Control for the Effects of Previous States (state dependence)

- Much of human behaviour is influenced by previous behaviour and outcomes (positive feedback)
- McGinnis (1968) '*axiom of cumulative inertia*'

Improving Control for the Effects of Previous States (state dependence)

- Working in May = more likely to be working in June
- Married this year = more likely to be married next year
- Own your own house this quarter
- Travel to work by car this week

Improving Control for the Effects of Previous States (state dependence)

With panel data we may be able to include
past behaviour in the modelling process

Summary Message

There are methodological benefits...
but panel data are not a panacea!



Tweet - Longitudinal data enhance
our ability to investigate complicated
processes in the social world

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Sources of Longitudinal Data

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Analysing Repeated Cross-Sectional Surveys

- Often over-looked as a source of longitudinal information
- Many countries have cross-sectional surveys that are carried out on a regular basis
- They offer the possibility of pooling data for different years
- Not based on repeated contacts with the same individuals or households
- But offer opportunities to analyse general trends over time

National Survey of Health and Development: NSHD



Welcome to the 1958 National Child Development Study

Principal Investigator: **Prof Alissa Goodman**

The National Child Development Study (NCDS) follows the lives of more than 17,000 people born in England, Scotland and Wales in a single week of 1958. Over the course of cohort members' lives, the NCDS has collected information on health, educational development, economic circumstances, health behaviour, wellbeing, social participation and other factors.

The NCDS is managed by CLS and funded by the **Economic and Social Research Council**.

Welcome to the 1970 British Cohort Study

Principal Investigator: **Prof Alice Sullivan**

The 1970 British Cohort Study (BCS70) follows the lives of more than 17,000 people born in England, Scotland and Wales in a single week of 1970. Over the course of cohort members' lives, the BCS70 has collected information on health, physical, educational and social development, and economic circumstances among other factors.

The BCS70 is managed by CLS and funded by the **Economic and Social Research Council**.

Avon Longitudinal Study of Parents and Children



Growing Up in Scotland



History of Next Steps (LSYPE) in England

Next Steps, formerly known as the Longitudinal Study of Young People in England (LSYPE), is a major survey of people born between 1 September 1989 and 31 August 1990. Originally commissioned by the then Department for Education, Children, Schools and Families (now [Department for Education, DfE](#)), the study was designed to monitor the factors affecting educational progress, attainment and transitions following the end of compulsory schooling.

The study began in 2004 when the young people were aged 14-16. It followed them through their education, from independent schools, as well as Pupil Referral Units. The study has since followed the young people.

Youth Cohort Study

[Abstract](#) | [Access](#) | [Get started](#) | [FAQ](#) | [Related](#) | [Links](#) | [Search](#)

SERIES ABSTRACT

The Youth Cohort Study (YCS) began in 1985. It is a major programme of longitudinal research designed to monitor the behaviour and decisions of representative samples of young people aged 16 years onwards as they make the transition from compulsory education to further or higher education, or to the labour market. The YCS tries to identify and explain the factors which influence post-16 transitions, for example, educational attainment, training opportunities, experiences at school. To date the YCS covers 13 cohorts and over 40 surveys.

Administrative Data

- ONS – Longitudinal Study (England and Wales)
- Northern Ireland Longitudinal Study (NILS)
- Scottish Longitudinal Study
 - A panel study of 274k people based on Census records
 - <http://www.lscs.ac.uk/sls/>

Panel Dataset Examples (Household Panel Studies)

- US Panel Study of Income Dynamics (PSID)
 - began in 1968
 - <http://psidonline.isr.umich.edu/>
- Germany Socio-Economic Panel (SOEP)
 - began in 1984
 - <http://www.diw.de/en/soep>
- British Household Panel Survey BHPS
 - (1991 onwards)
 - 5k households, 10k adults, <http://www.iser.essex.ac.uk/survey/bhps>



BHPS | British Household Panel Survey

Home

The British Household Panel Survey began in 1991 and is a multi-purpose study whose unique value resides in the fact that:

- it follows the same representative sample of individuals - the panel - over a period of years;
- it is household-based, interviewing every adult member of sampled households;
- it contains sufficient cases for meaningful analysis of certain groups such as the elderly or lone parent families.

The wave 1 panel consists of some 5,500 households and 10,300 individuals drawn from 250 areas of Great Britain. Additional samples of 1,500 households in each of Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2,000 households was added in Northern Ireland, making the panel suitable for UK-wide research.

BHPS

[+ About](#)[Acquiring the data](#)[+ Documentation](#)[+ Scientific steering committee](#)[Quality profile](#)[+ Faqs](#)[Updates](#)[Publications](#)

Nuisance calls claiming to be 'British Household Survey'

We have recently received a

[Home](#) > [BHPS](#) > [Documentation](#) > [Volb](#)

BHPS Documentation - Volume B - the Codebook

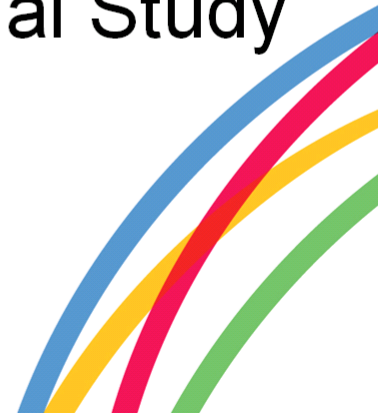
[Menu](#)

You may access the codebook material in several ways:

- by consulting the [Subject Category Thesaurus](#) in order to find suitable index term(s).
- by selecting a specific [Index Term](#)
- by viewing a list of BHPS [Record Types](#)
- by Wave:
 - [Wave One \(A\)](#)
 - [Wave Two \(B\)](#)
 - [Wave Three \(C\)](#)
 - [Wave Four \(D\)](#)
 - [Wave Five \(E\)](#)
 - [Wave Six \(F\)](#)
 - [Wave Seven \(G\)](#)
 - [Wave Eight \(H\)](#)
 - [Wave Nine \(I\)](#)



Understanding Society: the UK Household Longitudinal Study



<http://www.understandingsociety.org.uk/>

- Understanding Society (US)
 - Also known as the UK Household Longitudinal Study (UKHLS)
- Began in January 2009
- Incorporates and extends the BHPS
- 40k UK households (4K Scottish households)
- 4k households in a special ethnic minorities sample
- Innovations include:
 - Linking to administrative data; spatial data; biometric data; qualitative data; child data (from age 10)





















<http://www.understandingsociety.org.uk/>

Understanding Society Sample

- Approx. 27,000 households -
 - The fieldwork for this sample commenced in January 2009
- A boost ethnic minority sample,
 - focussed on five main ethnic minority groups, comprising 4,000 households
- Incorporating the BHPS sample of approximately 8,400 households
- An Innovation Panel of 1500 households to enable methodological research
 - (panel began in January 2008)

Understanding Society

- Focus on new research issues
- Opportunities for mixed methods:
 - Data linkage admin, organisation, spatial
 - Bio-markers and health indicators
 - Qualitative data
 - Other non-standard data: diaries, visual, audio

UKHLS 2016-18			UKHLS Wave 8 
UKHLS 2015-17			UKHLS Wave 7 
UKHLS 2014-16			UKHLS Wave 6 
UKHLS 2013-15			UKHLS Wave 5 
UKHLS 2012-14			UKHLS Wave 4 
UKHLS 2011-13			UKHLS Wave 3 
UKHLS 2010-12			UKHLS Wave 2 
UKHLS 2009-11			UKHLS Wave 1 
BHPS 2008			BHPS Wave 18 
BHPS 2007			BHPS Wave 17 
BHPS 2006			BHPS Wave 16 
BHPS 2005			BHPS Wave 15 
BHPS 2004			BHPS Wave 14 
BHPS 2003			BHPS Wave 13 
BHPS 2002			BHPS Wave 12 
BHPS 2001			BHPS Wave 11 
BHPS 2000			BHPS Wave 10 
BHPS 1999			BHPS Wave 9 
BHPS 1998			BHPS Wave 8 
BHPS 1997			BHPS Wave 7 

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Analysing Repeated Cross-Sectional Data

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Studying Longer Term Trends

- Many sources of 'repeated' cross-sectional data
- Rapid progress can be made
- Standard statistical approaches (e.g. regression models)
- Comparability (equivalence) is the central challenge
- How should time be represented

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Analysing Duration Data

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Alternative terminology

- Duration models
- Survival models
- Cox regression
- Cox models
- Failure time analysis
- Hazard models
- Event history analysis

Models for duration data allow the data analyst to assess the relative influence of a number of explanatory factors upon how long it takes for an event to occur

Original paper Cox (1972)

Applications

- Study the lifetimes of machine components in engineering
- Duration of unemployment in economics
- Time taken to complete cognitive tasks in psychology
- Lengths of tracks on a photographic plate in particle physics
- Survival times of patients in clinical trials

Research Examples

Heckman and Borjas (1980) used duration modelling approaches to study unemployment

Blossfeld and Hakim (1997) studied female part-time employment

Mulder and Smits (1999) investigated first time home ownership

Lillard et al. (1995) studied premarital cohabitation and subsequent marital dissolution

Research Examples

Kiernan and Mueller (1998) undertook an analysis of divorce using the BHPS and the NCDS

Boyle et al. (2008) examined union dissolution using the Austrian Family and Fertility Survey (FFS)

Chan and Halpin (2002) used BHPS to examine gender role attitudes and the domestic division of labour on divorce

Pevalin and Ermisch (2004) investigated mental health, union dissolution and re-partnering

Measuring a Duration

Three requirements for correctly determining a duration

1. A starting time must be unambiguously defined
2. Time must have a defined unit of measurement
3. The event must be clearly defined

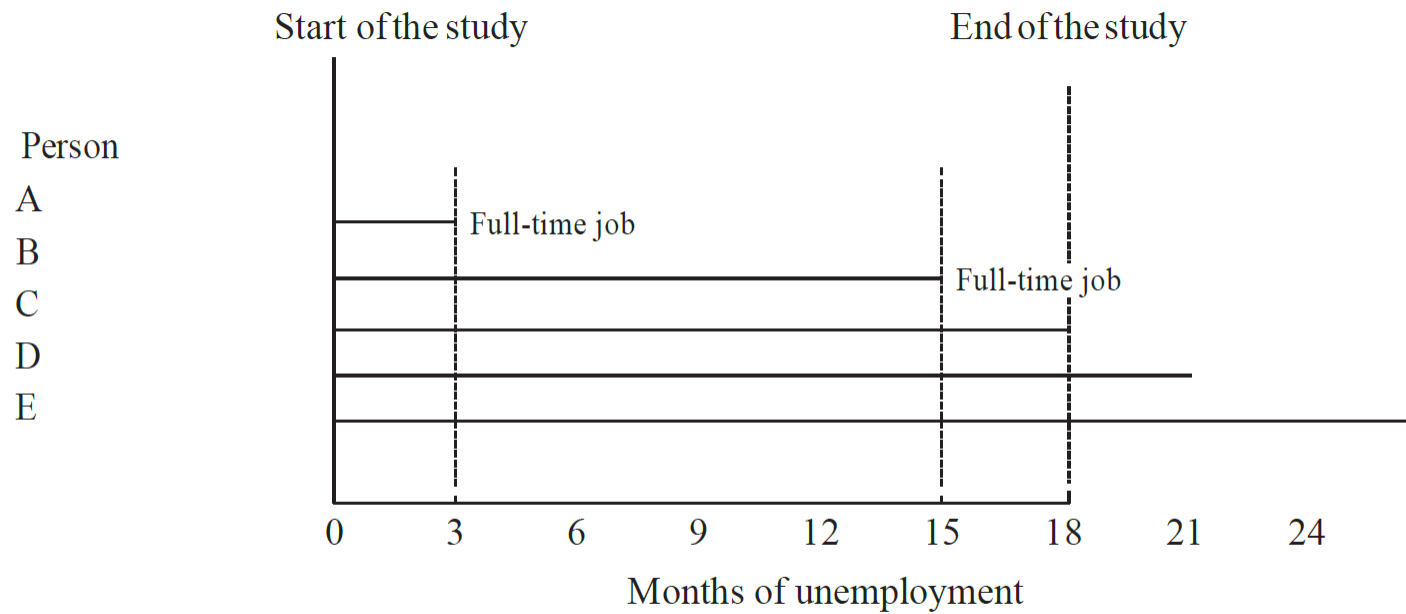


Figure 4 A diagram of a hypothetical study of unemployment

The Accelerated Life Model

Regression models can be estimated with duration data

Historically the log of the duration has been modelled

Censored Observations

- Censored observations affect regression model results
- The impact on the results on may sometimes be negligible
- Plewis (1997) states that when there is a very small proportion of censored cases they will have little effect, and an accelerated life model might still be suitable
- Supervisors, examiners and referees may not be convinced

Duration Modelling

- No longer directly modelling the duration
- The focus is on modelling the probability that an event occurs at time t , conditional on it not having occurred before t

Stata Compact Codebook for the College Skills Program Dataset

Variable	Obs	Unique	Mean	Min	Max	Label
id	628	628	314.5	1	628	student id
time	628	338	234.7038	2	1172	number of days until test passed
test	628	2	.8089172	0	1	test passed (or censored)
age	623	31	32.36918	20	56	age at enrolment
no_jobs	611	28	4.574468	0	40	number of previous jobs
mooc	628	2	.4904459	0	1	taught by massive open online course
campus	628	2	.2929936	0	1	college campus
quals1	628	2	.4601911	0	1	no qualifications
quals2	628	2	.1815287	0	1	lower qualifications (below A'level)
quals3	628	2	.3582803	0	1	higher qualifications(above A'level)

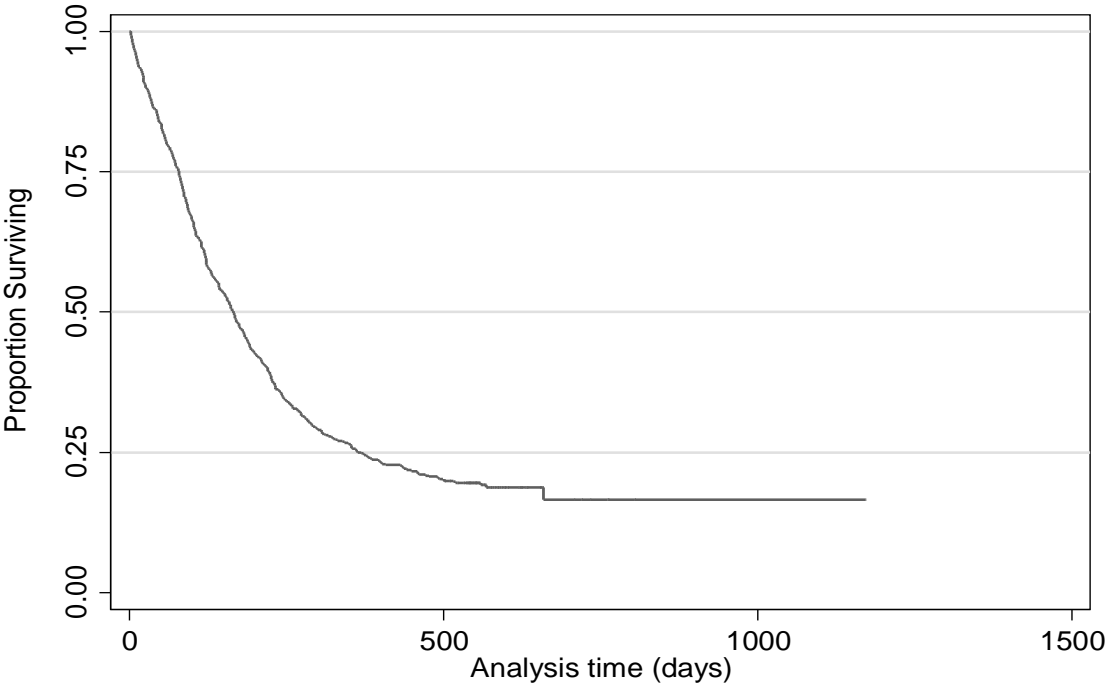
Stata Output: stdes Command for the College Skills Program Data

```
failure _d: test
analysis time _t: time
```

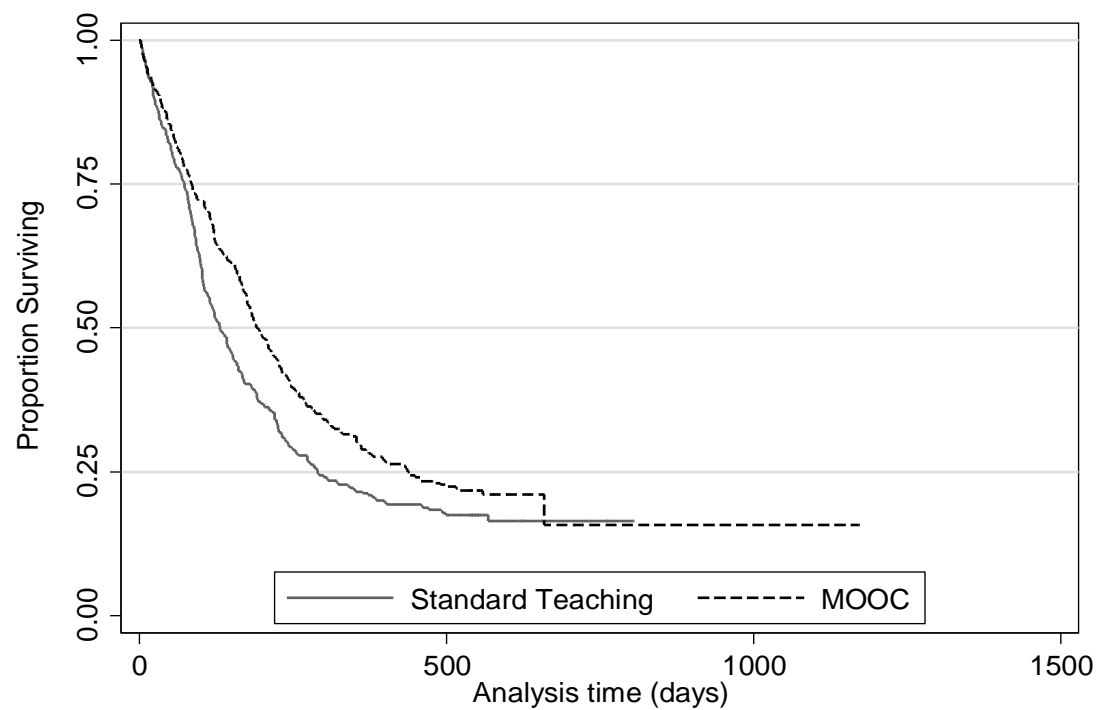
Category	total	----- per subject -----			
		mean	min	median	max

no. of subjects	628				
no. of records	628	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		234.7038	2	166	1172
subjects with gap	0				
time on gap if gap	0				
time at risk	147394	234.7038	2	166	1172
failures	508	.8089172	0	1	1

Stata Output: Kaplan-Meier Plot of Time to Passing the Test (College Skills Program Data)



Stata Output: Kaplan-Meier Plot of Time to Passing the Test (College Skills Program Data)



Stata Output: Log-Rank Test for Equality of Survivor Functions

```
failure _d:  test
analysis time _t:  time
```

Log-rank test for equality of survivor functions

	Events	Events
mooc	observed	expected
-----+-----		
0	265	235.80
1	243	272.20
-----+-----		
Total	508	508.00

```
chi2(1) =      6.80
Pr>chi2 =     0.0091
```

```
analysis time _t:  time
```

Stata Output: Cox Regression Model Time to Passing the Test (College Skills Program Data)

Cox regression -- Breslow method for ties

No. of subjects =	610	Number of obs =	610
No. of failures =	495		
Time at risk =	142994		
		LR chi2(6) =	34.94
Log likelihood =	-2851.0863	Prob > chi2 =	0.0000

	_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	age	-.0237543	.0075611	-3.14	0.002	-.0385737 -.0089349
	no_jobs	.034745	.0077538	4.48	0.000	.0195478 .0499422
	mooc	-.2540169	.091005	-2.79	0.005	-.4323834 -.0756504
	campus	-.1723881	.1020981	-1.69	0.091	-.3724966 .0277205
	quals2	.2467753	.1227597	2.01	0.044	.0061706 .4873799
	quals3	.125668	.1030729	1.22	0.223	-.0763513 .3276873

Stata Output: Test of the Effects of Previous Education in Cox Regression Model of Time to Passing the Test (College Skills Program Data)

(1) `quals2 = 0`

(2) `quals3 = 0`

`chi2(2) = 4.36`

`Prob > chi2 = 0.1130`

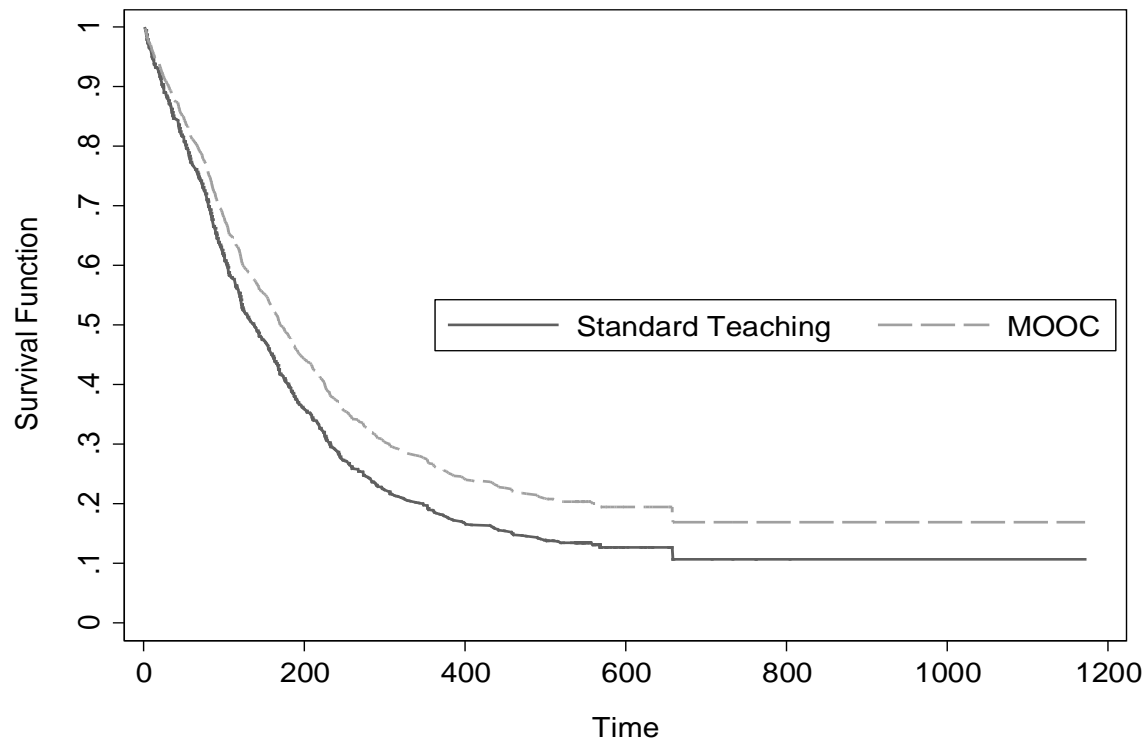
*Stata Output: Hazard Ratios Cox Regression Model Time to Passing the Test
(College Skills Program Data)*

Cox regression -- Breslow method for ties

No. of subjects =	610	Number of obs =	610
No. of failures =	495		
Time at risk =	142994		
		LR chi2(3) =	27.76
Log likelihood =	-2854.6735	Prob > chi2 =	0.0000

	_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	age	.9794475	.0072674	-2.80	0.005	.9653067 .9937955
	no_jobs	1.036128	.0078949	4.66	0.000	1.020769 1.051718
	mooc	.7940896	.0716076	-2.56	0.011	.6654445 .9476047

Stata Output: Time to Passing the Test - Survival Functions Comparing Women Aged 30 with 5 Previous Jobs by Teaching Methods (College Skills Program Data)



Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Analysing Panel Data (Part 1)

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

'The making of a causal inference is not a simple affair that can be reduced to a formula applied mechanically to a set of panel data on two or more variables'
(Duncan, 1972 p.36)

Wide format dataset – single survey contact

id	age	female	working_hours
001	20	0	37
002	30	1	39
003	40	1	45

Wide format dataset - repeated contacts

id	female	age_1971	age_1972	age_1973	work_hours_1971	work_hours_1972	work_hours_1973
001	0	20	21	22	37	40	35
002	1	30	31	32	39	40	35
003	1	40	41	42	45	45	15

Snapshot of long format dataset

id	year	age	hours	ln_wage
3	68	22	40	1.49
3	69	23	40	1.70
3	70	24	40	1.45

id: personal identification number

year: year of the survey

age: respondent's age in years

hours: number of hours per week normally worked in main job

ln_wage: log of weekly wages (adjusted for inflation)

Pooled Cross-Sectional Model

- Panel are pooled together and standard statistical model used (e.g. OLS)
- A good place to start to explore
- Results provide some initial information

Pooled Cross-Sectional Model

- Overall limitation of the pooled cross-sectional model is that it assumes that each observation (i.e. row within a long format dataset) is independent of other observations

Pooled Cross-Sectional Model

- With panel data we know that individual respondents contribute many times to the data (usually once per wave for many waves)
- Pooling all of the data violates the standard regression modelling assumption that each observation is independent

Pooled Cross-Sectional Model

- In practice standard errors that are too small
- Think about what this means for significance?

Robust Standard Errors

- Robust standard errors are sometimes known as Huber/White sandwich estimates of variance (see White, 1984, Huber, 1967)

id	year	age	hours	ln_wage
3	68	22	40	1.49
3	69	23	40	1.70
3	70	24	40	1.45

Collapsed dataset of the mean values

Id	year \bar{x}	age \bar{x}	hours \bar{x}	ln_wage \bar{x}
3	60	23	40	1.55

id: personal identification number

year \bar{x} : mean year of the survey

age \bar{x} : mean of respondent's age in years

hours \bar{x} : mean number of hours per week normally worked in main job

ln_wage \bar{x} : mean log of weekly wages (adjusted for inflation)

Between Effects Model

Estimates a standard cross-sectional model on the data

A regression model with Y the mean of the log of weekly wages (adjusted for inflation)

X vars

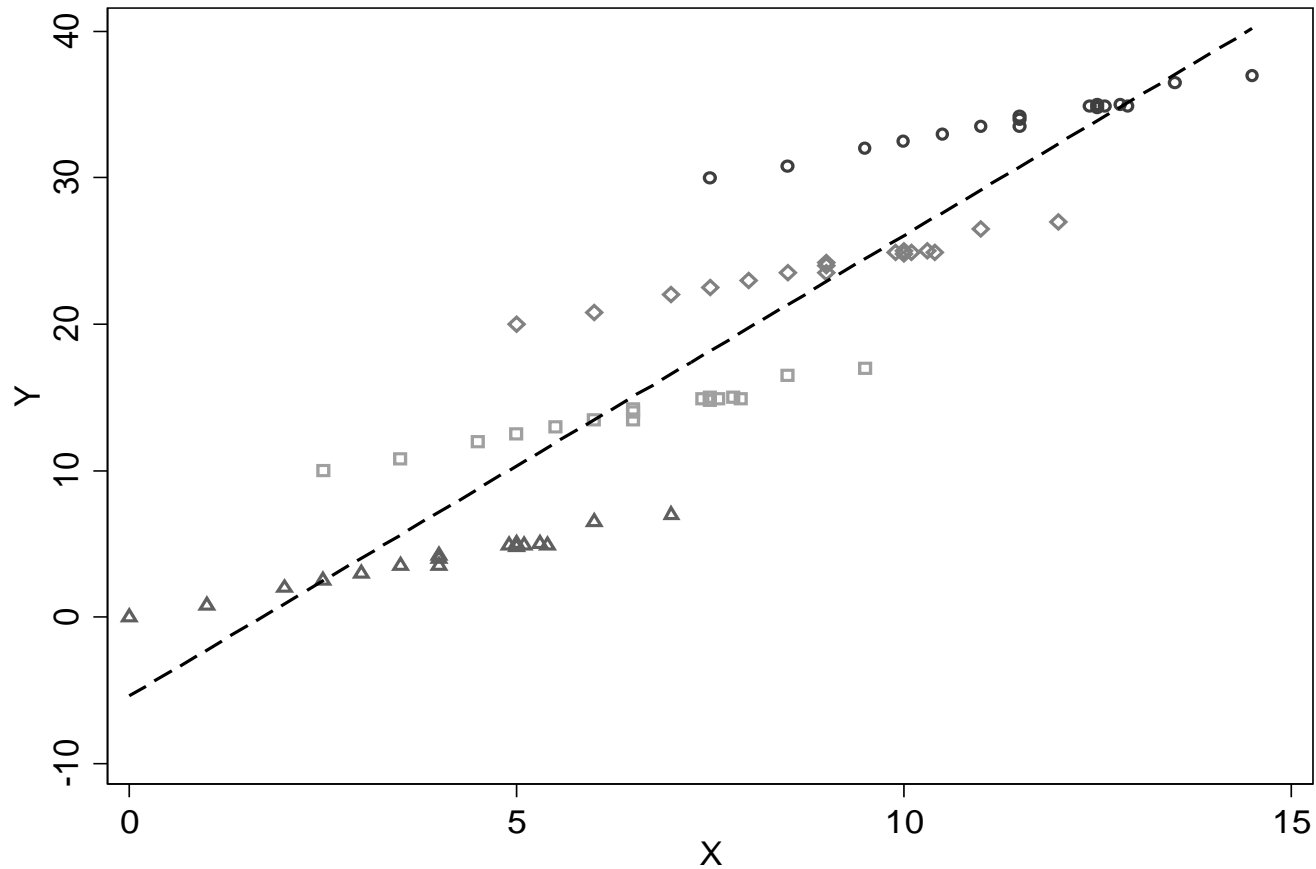
- mean hours per week normally worked in the respondent's main job
- mean age across the three waves of the survey

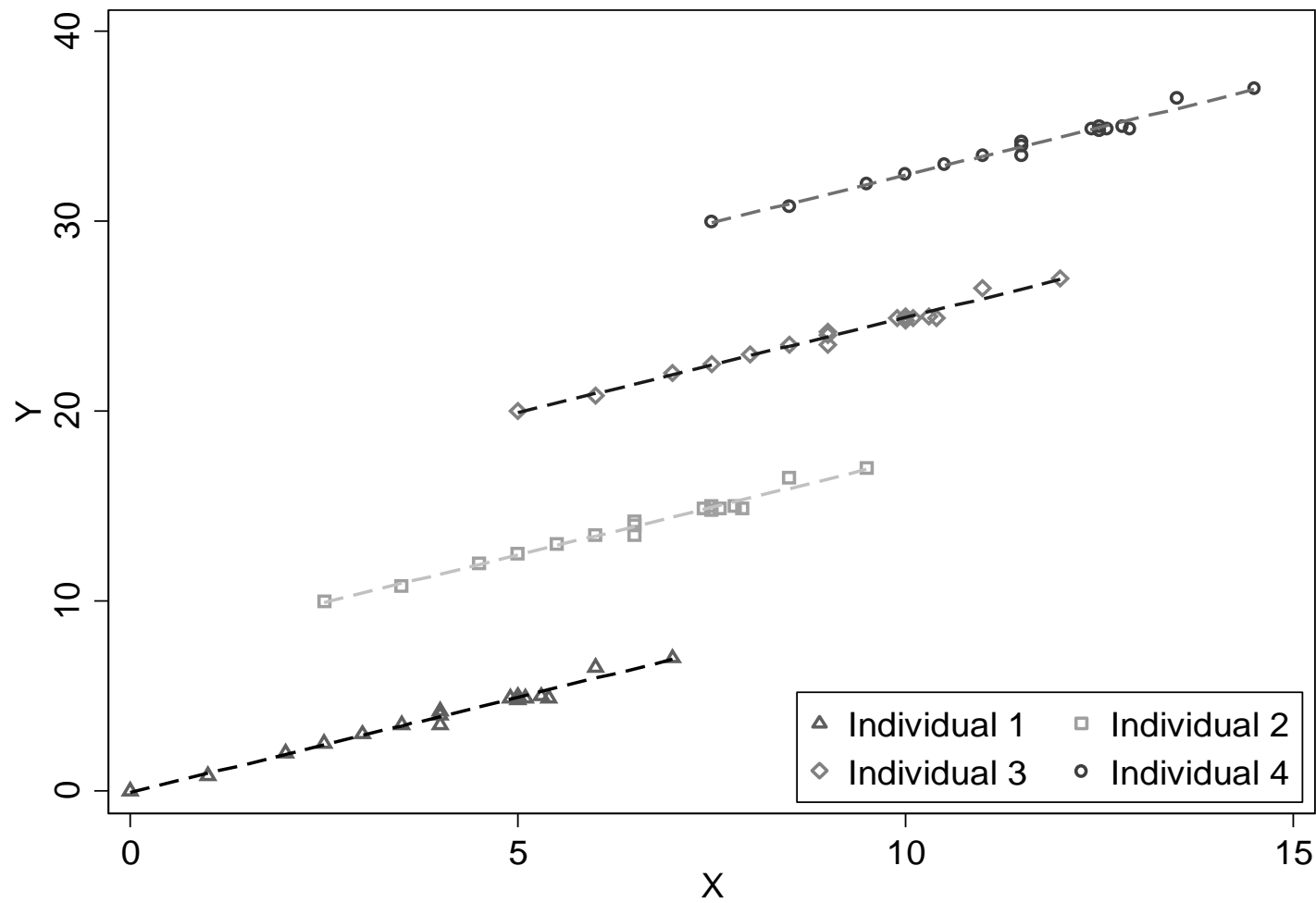
Between Effects Model

Because now there is only one row of data per respondent the problem of non-independence of observations in the original (long format) panel data is sidestepped

What might the limitation of this approach be?

A Thought Experiment...





The Fixed Effects Panel Model

- Concentrates on change over time within an individual respondent
- Can include explanatory variables that, for the individual respondent, change over time (e.g. age, monthly income and body mass index)
- In general cannot include explanatory variables that, for the individual respondent, are time-constant (e.g. town of birth, birth weight, father's occupation when respondent was aged 14)
- Has the potentially attractive property of providing robust estimates when observed explanatory variables are correlated with the unobserved effects

The Random Effects Panel Model

- Analyses both change within an individual respondent's outcomes, and differences between respondents' outcomes
- Can include explanatory variables that, for the individual respondent, change over time (e.g. age, monthly income and body mass index)
- Can include explanatory variables that, for the individual respondent, are time-constant (e.g. town of birth, birth weight, father's occupation when respondent was aged 14)
- Makes the assumption that observed explanatory variables are not correlated with the unobserved effects

Notation

Pooled Cross-Sectional Regression Model

$$(1) \quad Y_{it} = \beta_0 + \beta_1 X_{1it} + \dots + \beta_k X_{kit} + \varepsilon_{it}$$

Fixed Effects Panel Regression Model

$$(2) \quad Y_{it} = \beta_0 + \lambda_i + \beta_1 X_{1it} + \dots + \beta_k X_{kit} + \varepsilon_{it}$$

Random Effects Panel Regression ('random intercepts' version)

$$(3) \quad Y_{it} = \beta_0 + \beta_1 X_{1it} + \dots + \beta_k X_{kit} + \upsilon_i + \varepsilon_{it}$$

A Toy Example

Variable	Obs	Mean	Min	Max	Label
y	32	5.38	1	11	y outcome variable
id	32	4.50	1	8	id
female	32	.50	0	1	female
wave2	32	.25	0	1	wave 2
wave3	32	.25	0	1	wave 3
wave4	32	.25	0	1	wave 4

OLS Regression

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	.625	.4187448	1.49	0.147	-.2341934	1.484193
wave2	.75	.5921946	1.27	0.216	-.465083	1.965083
wave3	3.5	.5921946	5.91	0.000	2.284917	4.715083
wave4	6.25	.5921946	10.55	0.000	5.034917	7.465083
_cons	2.4375	.4681709	5.21	0.000	1.476893	3.398107

Between Effects Model

Between regression (regression on group means)	Number of obs	=	32
Group variable: id	Number of groups	=	8

R-sq:

Obs per group:

within	=	.	min	=	4
between	=	0.2500	avg	=	4.0
overall	=	0.0133	max	=	4

$$F(1, 6) = 2.00$$

sd(u_i + avg(e_i.))= .625 Prob F = 0.2070

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
female	.625	.4419417	1.41	0.207	-.4563925 1.706392
wave2	0	(omitted)			
wave3	0	(omitted)			
wave4	0	(omitted)			
_cons	5.0625	.3125	16.20	0.000	4.29784 5.82716

Fixed Effects Model

Fixed-effects (within) regression Number of obs = 32
Group variable: id Number of groups = 8

R-sq: Obs per group:
 within = 0.8722 min = 4
 between = . avg = 4.0
 overall = 0.8259 max = 4

corr(u_i, Xb) = -0.0000 F(3,21) = 47.77
 Prob F = 0.0000

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
female	0	(omitted)				
wave2	.75	.5824824	1.29	0.212	-.4613384	1.961338
wave3	3.5	.5824824	6.01	0.000	2.288662	4.711338
wave4	6.25	.5824824	10.73	0.000	5.038662	7.461338
_cons	2.75	.4118772	6.68	0.000	1.893454	3.606546
-----+-----						
sigma_u	.6681531					
sigma_e	1.1649647					
rho	.24752475	(fraction of variance due to u_i)				

F test that all u_i=0: F(7, 21) = 1.32 Prob F = 0.2914

Random Effects Model

```
Random-effects GLS regression              Number of obs   =        32
Group variable: id                        Number of groups  =         8

R-sq:                                     Obs per group:
    within = 0.8722                               min =         4
    between = 0.2500                              avg  =        4.0
    overall = 0.8392                              max  =         4

                                           Wald chi2(4)      =       145.32
corr(u_i, X)  = 0 (assumed)              Prob chi2        =       0.0000

-----+-----
      y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
    female |      .625   .4419417     1.41   0.157    - .2411899     1.49119
    wave2  |      .75   .5824824     1.29   0.198    - .3916445     1.891644
    wave3  |     3.5   .5824824     6.01   0.000     2.358356     4.641644
    wave4  |     6.25   .5824824    10.73   0.000     5.108356     7.391644
    _cons  |     2.4375   .474224     5.14   0.000     1.508038     3.366962
-----+-----

    sigma_u |   .22658174
    sigma_e |   1.1649647
         rho |   .03645008   (fraction of variance due to u_i)
-----+-----
```

Comparing Models

	BE	FE	RE
	b(se)	b(se)	b(se)
female	0.625 (0.442)	0.000 (.)	0.625 (0.442)
wave2	0.000 (.)	0.750 (0.582)	0.750 (0.582)
wave3	0.000 (.)	3.500*** (0.582)	3.500*** (0.582)
wave4	0.000 (.)	6.250*** (0.582)	6.250*** (0.582)
_cons	5.063*** (0.313)	2.750*** (0.412)	2.438*** (0.474)
n	32	32	32
BE:	Between Effects Model		
FE:	Fixed Effects Panel Model		
RE:	Random Effects Panel Model		

Concluding Thoughts

Random effects panel model is using (or borrowing) some information from the fixed effects panel model

At the same time it is borrowing some information from the between effects model

This should illustrate why econometricians often make oral statements such as 'the random effects panel model is a matrix weighted average of the within-effects (fixed effects) and the between effects'

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Analysing Panel Data (Part 2)

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Example: A BHPS Panel Data File

Example: A BHPS Panel Data File

first 10 years;
25 - 35 year olds;
'essex originals';
males;
working full-time;

yvar is wPAYNU2 Usual net pay per month: current
job (now adjusted for inflation)

Variable	Obs	Unique	Mean	Min	Max	Label
pid	8412	1324	1.99e+07	1.00e+07	1.07e+08	cross-wave person identifier
wave	8412	10	5.382668	1	10	wave of the BHPS
zhid	8412	8332	5794920	1000381	1.07e+07	household identification number
zpdo	8412	7	1.497147	1	7	person number
zdoby	8412	9	1960.981	1957	1965	year of birth
zpaynu2	6098	3898	1050.62	50.14124	9741.704	(deflated 1991)usual net pay per month...
zjbhrs	6435	60	40.47677	0	99	no. of hours normally worked per week
zjbcssm	7503	558	34.75936	.56	90.32	cambridge scale males: present job
pacssm	6827	347	30.12809	.56	85.04	cambridge scale males : father's job
graduate	7924	2	.1680969	0	1	Graduates (zqfachi)
zregage	8412	19	9.178673	0	18	age at interview-25

Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
pid	8,412	19933161	15557954	10004531	107127271
wave	8,412	5.382668	2.883692	1	10
zhid	8,412	5794920	2814730	1000381	10677259
zpno	8,412	1.497147	.8551585	1	7
zdoby	8,412	1960.981	2.607186	1957	1965
zpaynu2	6,098	1050.62	488.2907	50.14124	9741.704
zjbhrs	6,435	40.47677	7.895388	0	99
zjbcssm	7,503	34.75936	19.10313	.56	90.32
pacssm	6,827	30.12809	19.00986	.56	85.04
graduate	7,924	.1680969	.3739759	0	1
zregage	8,412	9.178673	3.853988	0	18

Pooled Cross-Sectional Model

```
. reg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav
```

Source		SS		df	MS	Number of obs	=	5,097
-----+-----						F(14, 5082)	=	128.85
Model		325506963		14	23250497.3	Prob > F	=	0.0000
Residual		917001747		5,082	180441.115	R-squared	=	0.2620
-----+-----						Adj R-squared	=	0.2599
Total		1.2425e+09		5,096	243820.391	Root MSE	=	424.78

Pooled Cross-Sectional Model (Continued)

zpaynu2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zjbhrs	9.559008	.8714275	10.97	0.000	7.850635	11.26738
zjbcssm	8.709623	.370681	23.50	0.000	7.982928	9.436317
pacssm	2.099019	.354503	5.92	0.000	1.40404	2.793998
graduate	195.6784	17.63807	11.09	0.000	161.1001	230.2566
zregage	6.51771	2.257654	2.89	0.004	2.091734	10.94368
wave						
2	32.02204	26.09955	1.23	0.220	-19.14433	83.1884
3	47.53042	26.53109	1.79	0.073	-4.481953	99.5428
4	32.50858	27.03488	1.20	0.229	-20.49144	85.5086
5	37.20393	27.6692	1.34	0.179	-17.03963	91.44749
6	83.98088	28.31416	2.97	0.003	28.47293	139.4888
7	72.86194	28.92154	2.52	0.012	16.16326	129.5606
8	96.8642	29.82545	3.25	0.001	38.39346	155.3349
9	139.523	31.31379	4.46	0.000	78.13451	200.9116
10	138.2166	32.437	4.26	0.000	74.62611	201.8071
_cons	135.0271	43.56656	3.10	0.002	49.61787	220.4363

Pooled Cross-Sectional Model

With robust standard errors

```
. reg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav, cluster(pid)
```

Linear regression

Number of obs	=	5,097
F(14, 824)	=	23.09
Prob > F	=	0.0000
R-squared	=	0.2620
Root MSE	=	424.78

(Std. Err. adjusted for 825 clusters in pid)

Pooled Cross-Sectional Model

With robust standard errors

zpaynu2	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zjbhrs	9.559008	1.943735	4.92	0.000	5.743753	13.37426
zjbcssm	8.709623	.7934006	10.98	0.000	7.152299	10.26695
pacssm	2.099019	1.115752	1.88	0.060	-.0910308	4.289069
graduate	195.6784	59.72225	3.28	0.001	78.45271	312.904
zregage	6.51771	4.515835	1.44	0.149	-2.346185	15.3816
wave						
2	32.02204	13.90313	2.30	0.022	4.732323	59.31175
3	47.53042	18.75842	2.53	0.011	10.71052	84.35033
4	32.50858	20.47166	1.59	0.113	-7.674154	72.69131
5	37.20393	25.89333	1.44	0.151	-13.62072	88.02857
6	83.98088	30.03699	2.80	0.005	25.02286	142.9389
7	72.86194	34.50542	2.11	0.035	5.133077	140.5908
8	96.8642	37.73043	2.57	0.010	22.80514	170.9233
9	139.523	43.37522	3.22	0.001	54.3841	224.662
10	138.2166	46.56261	2.97	0.003	46.82133	229.6119
_cons	135.0271	76.33256	1.77	0.077	-14.80204	284.8562

The Fixed Effects Panel Model

- Concentrates on change over time within an individual respondent
- Can include explanatory variables that, for the individual respondent, change over time (e.g. age, monthly income and body mass index)
- In general cannot include explanatory variables that, for the individual respondent, are time-constant (e.g. town of birth, birth weight, father's occupation when respondent was aged 14)
- Has the potentially attractive property of providing robust estimates when observed explanatory variables are correlated with the unobserved effects

The Random Effects Panel Model

- Analyses both change within an individual respondent's outcomes, and differences between respondents' outcomes
- Can include explanatory variables that, for the individual respondent, change over time (e.g. age, monthly income and body mass index)
- Can include explanatory variables that, for the individual respondent, are time-constant (e.g. town of birth, birth weight, father's occupation when respondent was aged 14)
- Makes the assumption that observed explanatory variables are not correlated with the unobserved effects

Notation

Pooled Cross-Sectional Regression Model

$$(1) \quad Y_{it} = \beta_0 + \beta_1 X_{1it} + \dots + \beta_k X_{kit} + \varepsilon_{it}$$

Fixed Effects Panel Regression Model

$$(2) \quad Y_{it} = \beta_0 + \lambda_i + \beta_1 X_{1it} + \dots + \beta_k X_{kit} + \varepsilon_{it}$$

Random Effects Panel Regression ('random intercepts' version)

$$(3) \quad Y_{it} = \beta_0 + \beta_1 X_{1it} + \dots + \beta_k X_{kit} + u_i + \varepsilon_{it}$$

```
. xtreg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav, fe
note: pacssm omitted because of collinearity
```

Fixed-effects (within) regression
Group variable: pid

Number of obs = 5,097
Number of groups = 825

R-sq:

within = 0.0973
between = 0.0047
overall = 0.0151

Obs per group:

min = 1
avg = 6.2
max = 10

corr(u_i, Xb) = -0.0716

F(13,4259) = 35.30
Prob > F = 0.0000

```
. xtreg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav, fe
```

```
note: pacssm omitted because of collinearity
```

```
Fixed-effects (within) regression      Number of obs   =      5,097
```

```
Group variable: pid                   Number of groups =      825
```

```
R-sq:
```

```
Obs per group:
```

```
    within  = 0.0973                      min =          1
```

```
    between = 0.0047                      avg  =         6.2
```

```
    overall = 0.0151                      max  =         10
```

```
F(13,4259)      =      35.30
```

```
corr(u_i, Xb)   = -0.0716                Prob > F      =      0.0000
```

```
. xtreg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav, fe
note: pacssm omitted because of collinearity
```

```
Fixed-effects (within) regression
Group variable: pid
```

```
Number of obs   =    5,097
Number of groups =    825
```

```
R-sq:
```

```
within  = 0.0973
between = 0.0047
overall = 0.0151
```

```
Obs per group:
```

```
min =    1
avg  =    6.2
max  =   10
```

```
F(13,4259)      =    35.30
```

```
Prob > F        =    0.0000
```

```
corr(u_i, Xb)   = -0.0716
```

```
. xtreg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav, fe
note: pacssm omitted because of collinearity
```

```
Fixed-effects (within) regression
Group variable: pid
```

```
Number of obs   =    5,097
Number of groups =    825
```

```
R-sq:
```

```
within  = 0.0973
between = 0.0047
overall = 0.0151
```

```
Obs per group:
```

```
min =    1
avg  =    6.2
max  =   10
```

```
corr(u_i, Xb) = -0.0716
```

```
F(13,4259)      =    35.30
Prob > F        =    0.0000
```


zpaynu2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zjbhrs	3.162216	.7598803	4.16	0.000	1.672455	4.651978
zjbcssm	.7083871	.4688331	1.51	0.131	-.2107702	1.627544
pacssm	0	(omitted)				
graduate	-68.42179	58.82074	-1.16	0.245	-183.7411	46.89752
zregage	9.779486	18.59062	0.53	0.599	-26.66781	46.22679
wave						
2	33.77626	23.69473	1.43	0.154	-12.67775	80.23027
3	50.54727	39.7453	1.27	0.204	-27.37422	128.4688
4	36.4279	57.69603	0.63	0.528	-76.68639	149.5422
5	38.07226	75.6673	0.50	0.615	-110.2751	186.4196
6	85.96321	93.47347	0.92	0.358	-97.2935	269.2199
7	87.99939	111.601	0.79	0.430	-130.7967	306.7955
8	115.7779	130.037	0.89	0.373	-139.1623	370.7181
9	161.5511	148.3231	1.09	0.276	-129.2394	452.3416
10	172.5317	167.0504	1.03	0.302	-154.9742	500.0375
_cons	751.7321	98.72638	7.61	0.000	558.177	945.2873

```

-----+-----
sigma_u | 436.52027
sigma_e | 251.17185
rho      | .75126958    (fraction of variance due to u_i)
-----
F test that all u_i=0: F(824, 4259) = 12.59          Prob > F = 0.0000

```

areg

```
. areg zpaynu2 zjbhrs zjbcssm pacssm graduate zregage i.wav, absorb(pid)
note: pacssm omitted because of collinearity
```

Linear regression, absorbing indicators	Number of obs	=	5,097
	F(13, 4259)	=	35.30
	Prob > F	=	0.0000
	R-squared	=	0.7838
	Adj R-squared	=	0.7413
	Root MSE	=	251.1718

zpaynu2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
zjbhrs	3.162216	.7598803	4.16	0.000	1.672455	4.651978
zjbcssm	.7083871	.4688331	1.51	0.131	-.2107702	1.627544
pacssm	0 (omitted)					
graduate	-68.42179	58.82074	-1.16	0.245	-183.7411	46.89752
zregage	9.779486	18.59062	0.53	0.599	-26.66781	46.22679
wave						
2	33.77626	23.69473	1.43	0.154	-12.67775	80.23027
3	50.54727	39.7453	1.27	0.204	-27.37422	128.4688
4	36.4279	57.69603	0.63	0.528	-76.68639	149.5422
5	38.07226	75.6673	0.50	0.615	-110.2751	186.4196
6	85.96321	93.47347	0.92	0.358	-97.2935	269.2199
7	87.99939	111.601	0.79	0.430	-130.7967	306.7955
8	115.7779	130.037	0.89	0.373	-139.1623	370.7181
9	161.5511	148.3231	1.09	0.276	-129.2394	452.3416
10	172.5317	167.0504	1.03	0.302	-154.9742	500.0375
_cons	751.7321	98.72638	7.61	0.000	558.177	945.2873
pid	F(824, 4259) =		12.593	0.000	(825 categories)	

xtreg , re (random effects)

```
. xtreg zpaynu2 zjbhr zjbcsm pacssm graduate zregage i.wav, re
```

Random-effects GLS regression

Group variable: pid

R-sq:

within = 0.0875

between = 0.2458

overall = 0.2291

corr(u_i, X) = 0 (assumed)

Number of obs	=	5,097
Number of groups	=	825

Obs per group:

min = 1

avg = 6.2

max = 10

Wald chi2(14) = 688.95

Prob > chi2 = 0.0000

xtreg , re (random effects)

```
. xtreg zpaynu2 zjbhr zjbcsm pacssm graduate zregage i.wav, re
```

Random-effects GLS regression
Group variable: pid

Number of obs	=	5,097
Number of groups	=	825

R-sq:

within	=	0.0875
between	=	0.2458
overall	=	0.2291

Obs per group:

min	=	1
avg	=	6.2
max	=	10

corr(u_i, X) = 0 (assumed)

Wald chi2(14)	=	688.95
Prob > chi2	=	0.0000

xtreg , re (random effects)

```
. xtreg zpaynu2 zjbhr zjbcsm pacssm graduate zregage i.wav, re
```

Random-effects GLS regression
Group variable: pid

Number of obs	=	5,097
Number of groups	=	825

R-sq:

within	=	0.0875
between	=	0.2458
overall	=	0.2291

Obs per group:

min	=	1
avg	=	6.2
max	=	10

corr(u_i, X) = 0 (assumed)

Wald chi2(14)	=	688.95
Prob > chi2	=	0.0000

xtreg , re (random effects)

```
. xtreg zpaynu2 zjbhr zjbcsm pacssm graduate zregage i.wav, re
```

Random-effects GLS regression
Group variable: pid

R-sq:

within = 0.0875
between = 0.2458
overall = 0.2291

corr(u_i, X) = 0 (assumed)

Number of obs = 5,097
Number of groups = 825

Obs per group:

min = 1
avg = 6.2
max = 10

Wald chi2(14) = 688.95
Prob > chi2 = 0.0000

xtreg , re (random effects)

```
. xtreg zpaynu2 zjbhr zjbcsm pacssm graduate zregage i.wav, re
```

Random-effects GLS regression
Group variable: pid

Number of obs = 5,097
Number of groups = 825

R-sq:

within = 0.0875
between = 0.2458
overall = 0.2291

Obs per group:

min = 1
avg = 6.2
max = 10

corr(u_i, X) = 0 (assumed)

Wald chi2(14) = 688.95
Prob > chi2 = 0.0000

zpaynu2	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
zjbhrs	4.095316	.7257752	5.64	0.000	2.672823	5.51781
zjbcssm	3.190851	.4147334	7.69	0.000	2.377988	4.003713
pacssm	3.949381	.7204518	5.48	0.000	2.537321	5.36144
graduate	202.0455	30.88336	6.54	0.000	141.5152	262.5758
zregage	9.659352	4.59312	2.10	0.035	.6570023	18.6617
wave						
2	32.94832	16.59421	1.99	0.047	.4242625	65.47238
3	48.03321	18.53548	2.59	0.010	11.70434	84.36207
4	31.86777	21.25793	1.50	0.134	-9.797011	73.53255
5	31.99669	24.47987	1.31	0.191	-15.98297	79.97635
6	78.77641	27.94861	2.82	0.005	23.99814	133.5547
7	77.89485	31.66425	2.46	0.014	15.83407	139.9556
8	100.9134	35.62999	2.83	0.005	31.07988	170.7469
9	144.7248	39.838	3.63	0.000	66.64376	222.8058
10	153.3713	44.0223	3.48	0.000	67.08919	239.6534
_cons	441.8567	45.78108	9.65	0.000	352.1274	531.5859

Breusch and Pagan Lagrangian multiplier test for random effects

$$\text{zpaynu2}[\text{pid}, \text{t}] = \text{Xb} + \text{u}[\text{pid}] + \text{e}[\text{pid}, \text{t}]$$

Estimated results:

	Var	sd = sqrt(Var)
-----+-----		
zpaynu2	243820.4	493.7817
e	63087.3	251.1718
u	110157.1	331.8993

Test: Var(u) = 0

chibar2(01) = 6525.74
Prob > chibar2 = 0.0000

	OLS b/se	BE b/se	FE b/se	RE b/se
zjbhrs	9.6*** (0.9)	14.1*** (2.1)	3.2*** (0.8)	4.1*** (0.7)
zjbcssm	8.7*** (0.4)	10.4*** (0.9)	0.7 (0.5)	3.2*** (0.4)
pacssm	2.1*** (0.4)	0.6 (0.8)	0.0 (.)	3.9*** (0.7)
graduate	195.7*** (17.6)	175.5*** (40.0)	-68.4 (58.8)	202.0*** (30.9)
zregage	6.5** (2.3)	6.8 (4.7)	9.8 (18.6)	9.7* (4.6)
1.wave	0.0 (.)	0.0 (.)	0.0 (.)	0.0 (.)
2.wave	32.0 (26.1)	-161.7 (121.5)	33.8 (23.7)	32.9* (16.6)
3.wave	47.5 (26.5)	-25.2 (142.1)	50.5 (39.7)	48.0** (18.5)
4.wave	32.5 (27.0)	-121.5 (162.1)	36.4 (57.7)	31.9 (21.3)
5.wave	37.2 (27.7)	364.7 (203.8)	38.1 (75.7)	32.0 (24.5)
6.wave	84.0** (28.3)	146.1 (198.4)	86.0 (93.5)	78.8** (27.9)
7.wave	72.9* (28.9)	-90.3 (160.0)	88.0 (111.6)	77.9* (31.7)
8.wave	96.9** (29.8)	-118.1 (111.5)	115.8 (130.0)	100.9** (35.6)
9.wave	139.5*** (31.3)	108.0 (127.4)	161.6 (148.3)	144.7*** (39.8)
10.wave	138.2*** (32.4)	3.9 (84.4)	172.5 (167.1)	153.4*** (44.0)
_cons	135.0** (43.6)	-16.5 (104.6)	751.7*** (98.7)	441.9*** (45.8)
N	5097	5097	5097	5097

	OLS b/se	BE b/se	FE b/se	RE b/se
zjbhrs	9.6*** (0.9)	14.1*** (2.1)	3.2*** (0.8)	4.1*** (0.7)
zjbcssm	8.7*** (0.4)	10.4*** (0.9)	0.7 (0.5)	3.2*** (0.4)
pacssm	2.1*** (0.4)	0.6 (0.8)	0.0 (.)	3.9*** (0.7)
graduate	195.7*** (17.6)	175.5*** (40.0)	-68.4 (58.8)	202.0*** (30.9)
zregage	6.5** (2.3)	6.8 (4.7)	9.8 (18.6)	9.7* (4.6)
1.wave	0.0 (.)	0.0 (.)	0.0 (.)	0.0 (.)
2.wave	32.0 (26.1)	-161.7 (121.5)	33.8 (23.7)	32.9* (16.6)
3.wave	47.5 (26.5)	-25.2 (142.1)	50.5 (39.7)	48.0** (18.5)
4.wave	32.5 (27.0)	-121.5 (162.1)	36.4 (57.7)	31.9 (21.3)
5.wave	37.2 (27.7)	364.7 (203.8)	38.1 (75.7)	32.0 (24.5)
6.wave	84.0** (28.3)	146.1 (198.4)	86.0 (93.5)	78.8** (27.9)
7.wave	72.9* (28.9)	-90.3 (160.0)	88.0 (111.6)	77.9* (31.7)
8.wave	96.9** (29.8)	-118.1 (111.5)	115.8 (130.0)	100.9** (35.6)
9.wave	139.5*** (31.3)	108.0 (127.4)	161.6 (148.3)	144.7*** (39.8)
10.wave	138.2*** (32.4)	3.9 (84.4)	172.5 (167.1)	153.4*** (44.0)
_cons	135.0** (43.6)	-16.5 (104.6)	751.7*** (98.7)	441.9*** (45.8)
N	5097	5097	5097	5097

	OLS b/se	BE b/se	FE b/se	RE b/se
zjbhrs	9.6*** (0.9)	14.1*** (2.1)	3.2*** (0.8)	4.1*** (0.7)
zjbcssm	8.7*** (0.4)	10.4*** (0.9)	0.7 (0.5)	3.2*** (0.4)
pacssm	2.1*** (0.4)	0.6 (0.8)	0.0 (.)	3.9*** (0.7)
graduate	195.7*** (17.6)	175.5*** (40.0)	-68.4 (58.8)	202.0*** (30.9)
zregage	6.5** (2.3)	6.8 (4.7)	9.8 (18.6)	9.7* (4.6)
1.wave	0.0 (.)	0.0 (.)	0.0 (.)	0.0 (.)
2.wave	32.0 (26.1)	-161.7 (121.5)	33.8 (23.7)	32.9* (16.6)
3.wave	47.5 (26.5)	-25.2 (142.1)	50.5 (39.7)	48.0** (18.5)
4.wave	32.5 (27.0)	-121.5 (162.1)	36.4 (57.7)	31.9 (21.3)
5.wave	37.2 (27.7)	364.7 (203.8)	38.1 (75.7)	32.0 (24.5)
6.wave	84.0** (28.3)	146.1 (198.4)	86.0 (93.5)	78.8** (27.9)
7.wave	72.9* (28.9)	-90.3 (160.0)	88.0 (111.6)	77.9* (31.7)
8.wave	96.9** (29.8)	-118.1 (111.5)	115.8 (130.0)	100.9** (35.6)
9.wave	139.5*** (31.3)	108.0 (127.4)	161.6 (148.3)	144.7*** (39.8)
10.wave	138.2*** (32.4)	3.9 (84.4)	172.5 (167.1)	153.4*** (44.0)
_cons	135.0** (43.6)	-16.5 (104.6)	751.7*** (98.7)	441.9*** (45.8)
N	5097	5097	5097	5097

	OLS b/se	BE b/se	FE b/se	RE b/se
zjbhrs	9.6*** (0.9)	14.1*** (2.1)	3.2*** (0.8)	4.1*** (0.7)
zjbcssm	8.7*** (0.4)	10.4*** (0.9)	0.7 (0.5)	3.2*** (0.4)
pacssm	2.1*** (0.4)	0.6 (0.8)	0.0 (.)	3.9*** (0.7)
graduate	195.7*** (17.6)	175.5*** (40.0)	-68.4 (58.8)	202.0*** (30.9)
zregage	6.5** (2.3)	6.8 (4.7)	9.8 (18.6)	9.7* (4.6)
1.wave	0.0 (.)	0.0 (.)	0.0 (.)	0.0 (.)
2.wave	32.0 (26.1)	-161.7 (121.5)	33.8 (23.7)	32.9* (16.6)
3.wave	47.5 (26.5)	-25.2 (142.1)	50.5 (39.7)	48.0** (18.5)
4.wave	32.5 (27.0)	-121.5 (162.1)	36.4 (57.7)	31.9 (21.3)
5.wave	37.2 (27.7)	364.7 (203.8)	38.1 (75.7)	32.0 (24.5)
6.wave	84.0** (28.3)	146.1 (198.4)	86.0 (93.5)	78.8** (27.9)
7.wave	72.9* (28.9)	-90.3 (160.0)	88.0 (111.6)	77.9* (31.7)
8.wave	96.9** (29.8)	-118.1 (111.5)	115.8 (130.0)	100.9** (35.6)
9.wave	139.5*** (31.3)	108.0 (127.4)	161.6 (148.3)	144.7*** (39.8)
10.wave	138.2*** (32.4)	3.9 (84.4)	172.5 (167.1)	153.4*** (44.0)
_cons	135.0** (43.6)	-16.5 (104.6)	751.7*** (98.7)	441.9*** (45.8)
N	5097	5097	5097	5097

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

https://github.com/vernongayle/longitudinal_warwick

2018

© Vernon Gayle

AQMEN

Analysing Panel Data (Part 3)

Vernon Gayle

Professor of Sociology & Social Statistics
University of Edinburgh

vernongayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN

Which Model Should We Use?

Gelman and Hill (2007), two leading statisticians, comment that the statistical literature is full of confusing and contradictory advice

Searle, Casella and McCulloch (1992) assert that because of conflicting definitions, it is no surprise that clear answers to the question 'fixed or random effects?' are unusual

- Researchers routinely ask, ‘Should I choose a fixed effects panel model or a random effects panel model?’
- The answer depends on what the data analyst is attempting to model
- The fixed effects panel model focuses upon the within-subject change
- The random effects panel model is influenced by both within-subject and between-subject patterns
- Both have potential advantages and limitations

```
.hausman fe re
```

---- Coefficients ----				
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fe	re	Difference	S.E.
logq	.9192846	.9066805	.0126041	.0153877
logf	.4174918	.4227784	-.0052867	.0058583
lf	-1.070396	-1.064499	-.0058974	.0255088

b = consistent under Ho and Ha; obtained from xtreg
 B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
 = 2.12
 Prob>chi2 = 0.5469
 (V_b-V_B is not positive definite)

Which Model Should We Use?

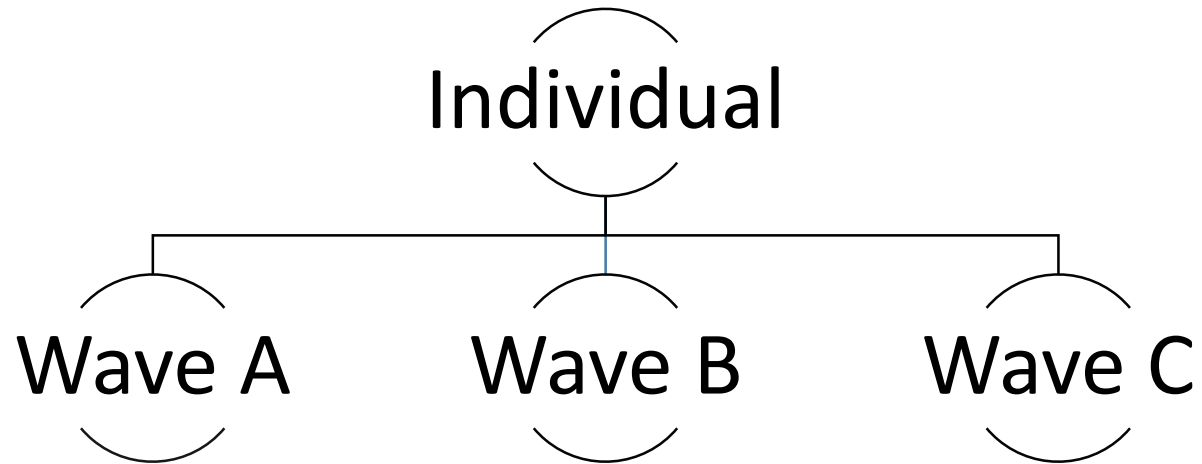
See Gayle and Lambert (2018)

Comparing fixed effects panel models and random effects panel models

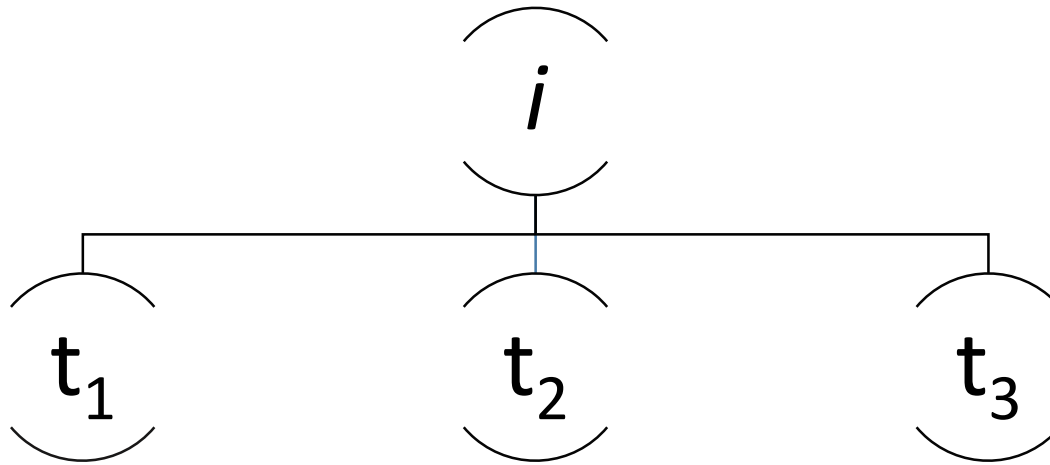
Alternative Literatures

- Econometrics
- Multilevel modelling (e.g. in the education literature)
- Epidemiology and biostatistics (e.g. public health)

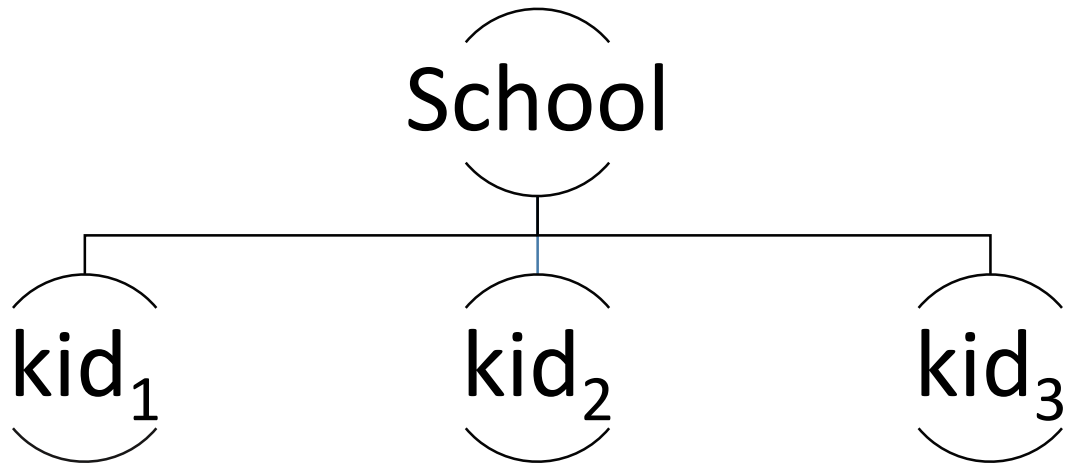
Panel Data Structure



Panel Data Structure



Hierarchical Data Structure



Stata Code

Panel Model

```
xtreg zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav, re mle
```

Multilevel Model

```
xtmixed zpaynu2 zjbhr zjbcssm pacssm graduate zregage i.wav|| pid:, mle
```

	(1)	(2)
	Panel RE	Multilevel
zpaynu2		
zjbhrs	3.983*** (0.723)	3.983*** (0.721)
zjbcssm	2.975*** (0.425)	2.975*** (0.415)
pacssm	4.075*** (0.759)	4.075*** (0.757)
graduate	195.5*** (32.05)	195.5*** (31.91)
zregage	9.742* (4.820)	9.742* (4.820)

_cons	449.7*** (46.94)	449.7*** (46.84)

sigma_u		
_cons	358.2*** (10.18)	

sigma_e		
_cons	251.9*** (2.734)	

lnsl_1_1		
_cons		5.881*** (0.0284)

lnsig_e		
_cons		5.529*** (0.0109)

N	5097	5097

Standard errors in parentheses

* p<0.05, ** p<0.01, *** p<0.001

Stata logs the sigma_u and sigma_e
in the standard xtmixed output

Mundlak-Chamberlain

In a nutshell...

The inclusion of means of the time-varying covariates within the random effects model

Table 6 Coefficients (β) and their standard errors (se) from the fixed effects panel model, the random effects panel model with Mundlak adjustment, and the random effects panel model – log wages

	(1) Fixed effects β_{fe} (s.e.)		(2) Random effects with Mundlak β_{mre} (s.e.)		(3) Random effects β_{re} (s.e.)	
Ft work	0.097	***	0.097	***	0.057	***
Experience(years)	(0.001)		(0.001)		(0.001)	
Weeks worked	0.001	*	0.001	*	0.002	**
	(0.001)		(0.001)		(0.001)	
Blue-collar occupation	-0.021		-0.021		-0.108	***
	(0.014)		(0.014)		(0.016)	
Individual's mean ft work experience (years)			-0.090	***		
			(0.002)			
Individual's mean weeks worked			0.010	**		
			(0.004)			
Individual's mean blue-collar occupation			-0.316	***		
			(0.034)			
Constant	4.709	***	6.164	***	5.523	***
	(0.038)		(0.212)		(0.047)	
n	4165		4165		4165	

Table 7 Coefficients (β) and their standard errors (se) from the fixed effects panel model, the random effects panel model with Mundlak adjustment, and the random effects panel model – Log total cost (per \$1000)

	(1) Fixed effects β_{fe} (s.e.)		(2) Random effects with Mundlak β_{mre} (s.e.)		(3) Random effects β_{re} (s.e.)	
Log output revenue index (passenger miles)	0.919 (0.030)	***	0.919 (0.030)	***	0.907 (0.026)	***
Log price of fuel	0.417 (0.015)	***	0.417 (0.015)	***	0.423 (0.014)	***
Load factor (average capacity of the fleet)	-1.070 (0.202)	***	-1.070 (0.202)	***	-1.064 (0.200)	***
Airline mean log output revenue index (passenger miles)			-0.137 (0.113)			
Airline mean log price of fuel			-5.941 (4.479)			
Airline mean Load factor (average capacity of the fleet)			-0.681 (2.751)			
Constant	9.714 (0.230)	***	85.808 (56.482)		9.628 (0.210)	***
n	90		90		90	

Data from Greene (1999).

Other Panel Models

Binary Outcomes

xtlogit	example in the .do file
xtprobit	example in the .do file
clogit	example in the .do file

Ordinal Outcomes

xtologit	random-effects ordered logistic models
xtoprobit	random-effects ordered probit models

Count Data

xtpoisson	panel data poisson models
xtnbreg	panel data negative binomial models

Dynamic Models

- Dynamic panel models extend panel models
- Appeal to the idea of using panel data to better understand 'state dependence'
- Lagged dependent variables as X vars
- Complicated because the lagged dependent variables will themselves be influenced by unobserved effects

Dynamic Models

- Standard panel estimation procedures will be inconsistent with lagged dependent variables
- Arellano and Bond (1991) derived a suitable estimator which is available using the Stata command *xtabond*
- Stewart (2006) *redprob*

Further Topics to Consider...

- The estimation and interpretation of interaction effects in statistical models (see Ai and Norton, 2003; Norton, Wang and Ai, 2004; Mitchell and Chen, 2005)
- Post-estimation measures and model evaluation (see Long and Freese, 2014)
- Missing data (see Carpenter and Kenward, 2012)
- Sample attrition, panel conditioning, interviewer effects and data collection modes (see Lynn, 2009)

Final Comment...

Angrist and Pischke (2008) playfully remarked that if applied research was easy then theorists would do it!

They also reassure readers that applied research is not as hard as the dense pages of Econometrica might lead us to believe

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh

vernon.gayle@ed.ac.uk
@Profbigvern

<https://github.com/vernongayle/>

2018

© Vernon Gayle

AQMEN