Contents

Longitudinal Data Analysis

Longitudinal data offer considerable statistical and analytical advantages to the social science researcher, including the ability to examine micro-level change (and stability), determine temporal ordering of events, and improved control for residual heterogeneity.

This book contains the materials underpinning a one-day course on *Longitudinal Data Analysis for Social Scientists* run by [Dr Diarmuid McDonnell](), UK Data Service. The course was first run on 2020-09-10.

Acknowledgements

We are grateful to the Bristol Doctoral College for supporting this training through its PGR Development Fund.

Further information

Please do not hesitate to get in contact if you have queries, criticisms and ideas regarding these materials: [Dr Diarmuid McDonnell]()

Essential Concepts

This section provides a concise overview of some important terms, concepts and analytcial approaches that are central to quantitative analyses of longitudinal data. It is aimed at people needing a quick refresher of key topics; the aim is **not** to teach you these topics for the first time.

Statistical modelling

Models help us make sense of the world and are more commonplace than you might think:

MacInnes (2017, p, 26) describes a model in more formal terms:

> A model is a simplified, often smaller-scale, version of reality; a summary statement that includes the essential aspects we are interested in and leaves out the extraneous detail…A good model focuses on what we want to investigate , and discards other features that are not relevant.

Statistical models are formal, numeric representations of a phenomenon and its explanatory factors, and are used both to understand and make predictions about said phenomenon.

For example, we can define a statistical model to predict whether somebody will finish their PhD as follows:

$$\text{Complete PhD} = \text{Receive funding} + \text{Good supervisors} + \text{Settled personal life} \tag{1.1}$$

Each of these factors contributes to the overall likelihood or chance of experiencing the outcome (completed PhD). Obviously this model ignores lots of other factors that are relevant to completing a PhD, but it's not a bad approximation and can be added to if we are in possession of more/better information on a PhD student.

Linear models

Equation 1.1 above is an example of a linear model whereby the outcome (completing a PhD) is a **linear** function of a set of explanatory factors.

Each explanatory factor has a distinct, linear effect on the outcome, and our prediction for the outcome is arrived at by adding together each of these effects.

Using equation 1.1, let's predict the probability of completing a PhD:

$$.8 = .4 + .2 + .2$$

In this fictional example, there is an 80% chance of completing a PhD if you receive funding (40%), have good supervisors (20%), and have a settled personal life (20%). Each factor contributes to the prediction but it is clear receiving funding is the most important factor.

Linear regression models

How do we assign values to the explanatory factors in a linear model? That's where linear regression comes in. *Linear regression* is known as the "workhorse" of quantitative social science (MacInnes, 2017) and for very good reason: many social phenomena can be modelled as a linear function of explanatory factors.

The linear regression equation (model) looks very similar to equation 1.1, just with some additional terms (parameters):

$$\mathrm{Y} = \alpha + \beta X + \epsilon \tag{1.2}$$

Where:

Y is a numeric outcome

α is a constant effect (think of this like an initial / baseline prediction of *Y* before we consider the effects of the explanatory factors)

X is a set of explanatory variables (factors that are included in the model)

β is the numeric estimate of the effect of the explanatory variables on the outcome

ϵ is an error term (residual), which captures the part of the outcome we cannot explain / predict using our explanatory variables and the constant term

Estimating and interpreting linear regression models

Regression is best understood by digging into some examples, so let's do that using the (in)famous `auto.dta` data set in Stata.

```
sysuse auto, clear
desc, f
```

```
(1978 Automobile Data)
```

```
Contains data from C:\Program Files (x86)\Stata14\ado\base/a/auto.dta
```

```
  obs:            74                              1978 Automobile Data
```

```
 vars:            12                          13 Apr 2014 17:45
 size:         3,182                          (_dta has notes)
--------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
--------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                Price
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g      origin     Car type
--------------------------------------------------------------------------------
Sorted by: foreign
```

We have 74 observations and 12 variables relating to the auto repair records for a set of cars. Let's say we want to understand the relationship between the price of a car (`price`) and its fuel efficiency (`mpg`) and mass (`weight`). We can state this statistical model using a slightly altered version of the general regression equation (1.2):

$$\mathrm{y}_i = \alpha + \beta 1 x_{1i} + \beta 2 x_{2i} + \epsilon_i \tag{1.3}$$

$$\epsilon_i = \mathrm{y}_i - \hat{y}_i \tag{1.4}$$

$$\hat{y}_i = \alpha + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \tag{1.5}$$

Now let's estimate this statistical model using linear regression:

```
regress price mpg weight
```

```
      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(2, 71)        =     14.74
```

```
       Model |   186321280         2  93160639.9   Prob > F        =    0.0000

    Residual |   448744116        71  6320339.67   R-squared       =    0.2934

-------------+----------------------------------   Adj R-squared   =    0.2735

       Total |   635065396        73  8699525.97   Root MSE        =      2514

------------------------------------------------------------------------------

       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

-------------+----------------------------------------------------------------

         mpg |  -49.51222   86.15604    -0.57   0.567    -221.3025     122.278

      weight |   1.746559   .6413538     2.72   0.008      .467736    3.025382

       _cons |   1946.069    3597.05     0.54   0.590    -5226.245    9118.382

------------------------------------------------------------------------------
```

How do we interpret the results produced by the linear regression model?

Let's start with the *coefficients* (effects) of the explanatory variables:

- For every one-unit increase in the fuel efficiency of a car, we predict the price of a car to decline by 50 dollars on average.
- For every one-unit increase in the weight of a car, we predict the price of a car to increase by 2 dollars on average.

The constant (`_cons`) represents our estimate of the price of a car if both `mpg` and `weight` are zero (obviously a nonsensical scenario).

How confident are we in the estimates of these effects?

- We fail to reject the null hypothesis that the coefficient of `mpg` is equal to zero (*statistically insignificant* as `P>|t| > .05`).
- We reject the null hypothesis that the coefficient of `weight` is equal to zero (*statistically significant* as `P>|t| < .05`).

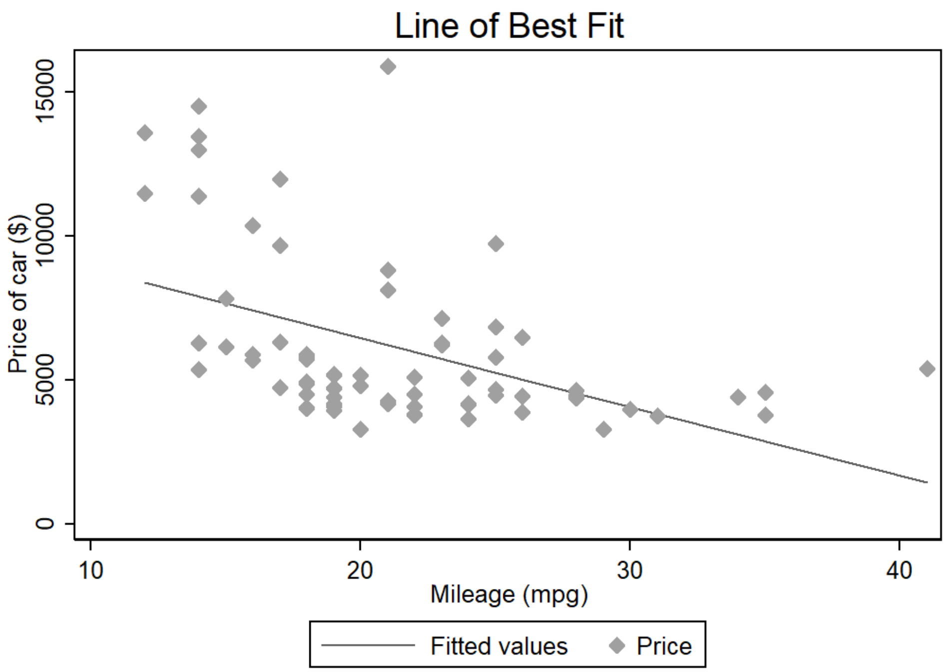
How good is the model overall at predicting the outcome?

- The explanatory variables are highly likely to have a non-zero effect on the outcome (`Prob > F = 0.0000`).
- The proportion of variance explained (`R-squared`) is 30%, suggesting that this model accounts for about one third of the variation in the price of a car. That is, price varies across cars and we can explain some degree of variation using this statistical model.

How does linear regression work?

Linear regression estimates coefficients for each of the explanatory variables using the **ordinary least squares (OLS)** estimator.

OLS selects the estimates ($\hat{\beta}_1, \hat{\beta}_2$ etc) that minimise the sum of the squared residuals.



Assumptions underpinning regression

Validity: data map to the research question. Another way of putting this is that the model is properly specified: only and all relevant explanatory variables are included.

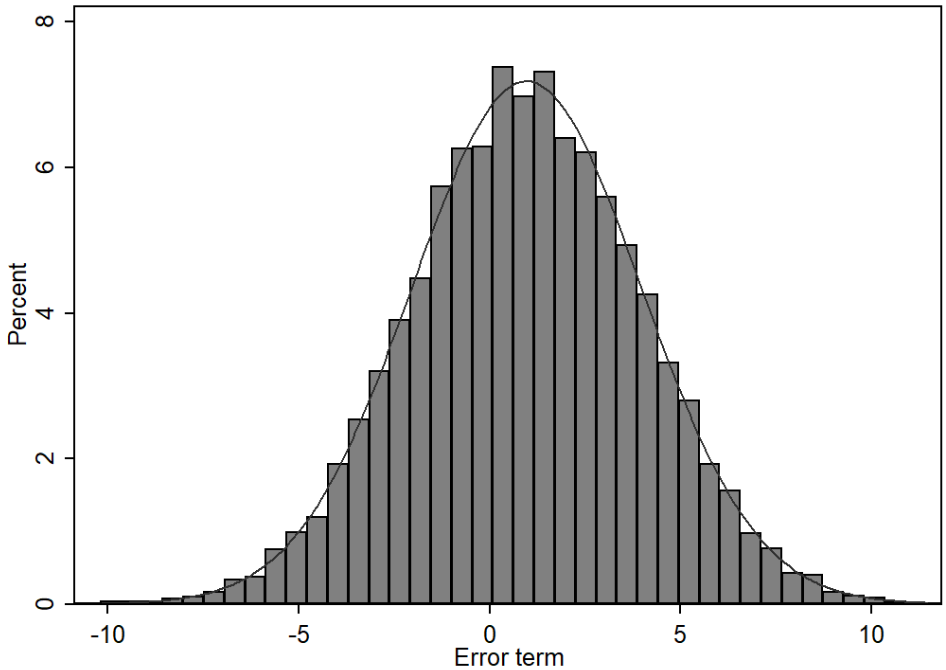
Additivity and linearity: the deterministic component of the model should be a linear function of the explanatory variables. The relationship between each explanatory variable and the outcome should be modelled in linear terms, and the predicted value of the outcome should be the sum of the coefficients for the explanatory variables (and constant).

Independence of errors: no correlation between the error term and the explanatory variables. If there is a correlation, it can lead to biased estimates.

Equal variance of errors: homoscedastic distribution of the errors. This means the degree to which your model is wrong is fairly constant across values of your explanatory variables.

Normality of errors: the errors follow a normal distribution (see below).

(Gelman and Hill, 2007)



If these assumptions are met, linear regression is considered **BLUE**:

- Best
- Linear
- Unbiased
- Estimator

Properties of estimators

Unbiased

We want our estimator to give the correct answer on average; that is, it can be wrong for individual applications of the estimator, but the average answer of these applications is correct (King et al., 1994).

$\mathrm{E}[\hat{\beta}] = \beta$

Consistent

We want our estimator to produce coefficients that converge to the true value as our sample size increases.

> A consistent estimator has the statistical property that as the number of data points increases it converges on the true value. (Gayle and Lambert, 2018, p. 89)

$\hat{\beta} \to \beta$ as $\mathrm{n} \to \mathrm{N}$.

Efficient

We want our estimator to be as precise as possible; that is, it minimises the variance of the estimate. An estimate, by definition, is uncertain and we would like to reduce that uncertainty to a minimum. Efficiency provides a way of distinguishing between unbiased estimators: An estimator that utilises more observations will be more efficient as it reduces the variance (King et al., 1994).

$\mathrm{Var}[\hat{\beta}]$ is minimised.

Introduction to Longitudinal Data

This section draws heavily on the work of Professor Vernon Gayle: Longitudinal Data Analysis for Social Scientists

What are longitudinal data?

At its simplest, longitudinal data contain a temporal dimension. This may be as simple as the data set containing variables that define the beginning and end of a social process (e.g., how long did somebody remain unemployed?). More often when we speak of longitudinal data we refer to data sets containing multiple observations of the same individuals.

Types of longitudinal study designs

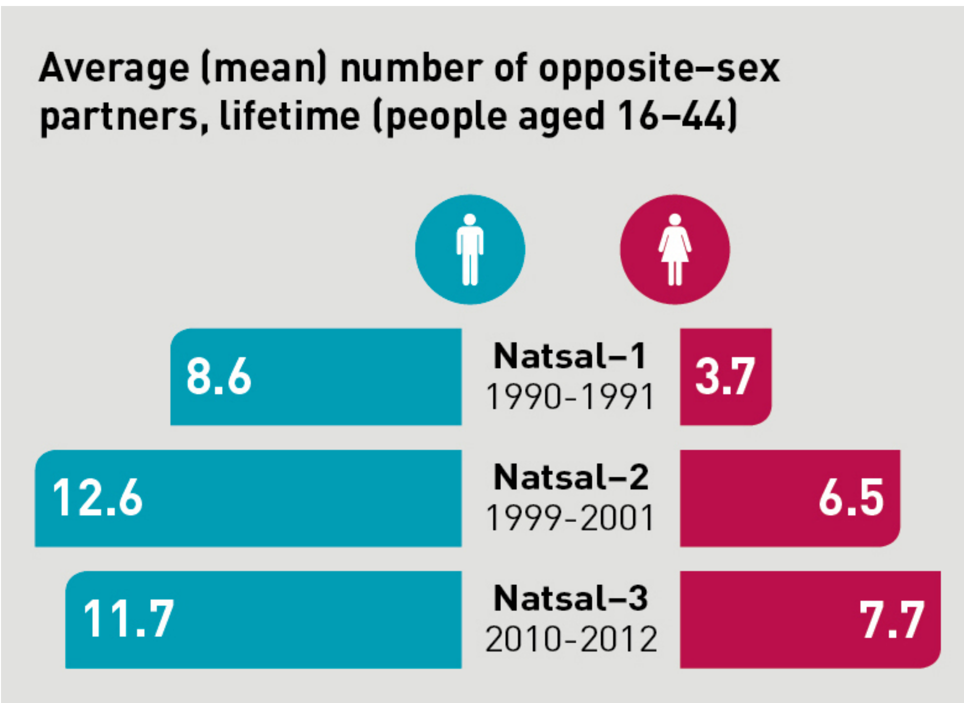
Repeated cross-sectional studies

Repeat samples of the same population over time:

- National Surveys of Sexual Attitudes and Lifestyles (NATSAL)
- British Social Attitudes Survey

Repeated cross-sectional studies allow analysis of change over time at the aggregate / macro level. For example, the mean number of opposite-sex sexual partners has increased over time in the UK for both men and women:

Figure 1.1.



Credit: Wellcome Trust/Paulo Estriga

Panel study

Groups of entities are repeatedly studied over time:

- UK Household Longitudinal Study (UKHLS)
- Panel Study of Income Dynamics (PSID)
- English Longitudinal Study of Aging (ELSA)

Panel studies collect data on the **same respondents** over time, and thus are known as *repeated contacts* data. For example, PSID has a module examining charitable giving of US households since 2000; this information is collected biennially and allows us to understand how the same households alter their giving behaviour over time (see figure 3.2 below).

Figure 1.2

Credit: [Changes to the Giving Landscape]

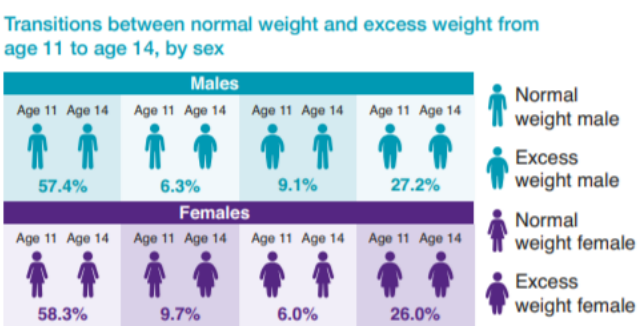
Cohort study

Following a particular group of entities over time:

- Millennium Cohort Study
- Growing Up in Scotland
- Whitehall Study II

The Millennium Cohort Study is a multi-wave survey of almost 20,000 children born in the UK during 2000/01, and is a representative sample of all children born during this period (Rafferty et al., 2015). It collects data at different periods (waves) of the children's lives, thus providing longitudinal information on the development and life histories of these children.

Figure 1.3.



Credit: [Child overweight and obesity: Initial findings from the Millennium Cohort Study Age 14 Survey]

Why use longitudinal data?

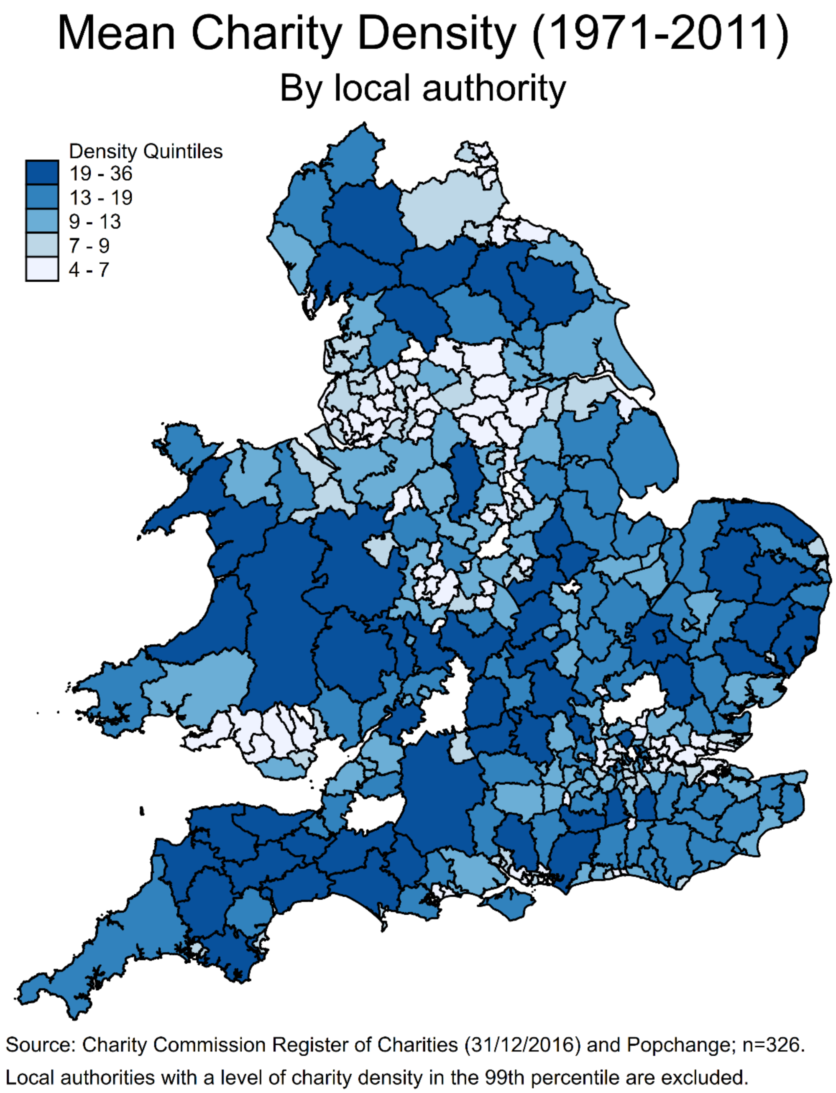
- UK has an unparalleled collection of longitudinal data resources.
- These resources are critical for analysing social change (and social stability).
- However they are costly to collect, clean and share, therefore strong justification needed.

Answering research questions

For many social science research projects cross-sectional data will be sufficient.

For example, if we are interested in understanding regional inequalities, it is sufficient to take a cross-section of data for these regions (e.g., a single census year) and describe variation in some measure of inequality. One of my recent research projects examined the distribution of charities across local authorities in England and Wales:

Figure 1.4.



The map displays the mean number of charities per 5000 residents across 326 local authorities in England and Wales. In essence I combined five census years to produce a cross-section of charity density between 1971 and 2011; that is I ignored the longitudinal component of my data and focused instead on making comparisons *between* local authorities.

Most social research projects can be improved by the analysis of longitudinal data.

Figure 1.5.

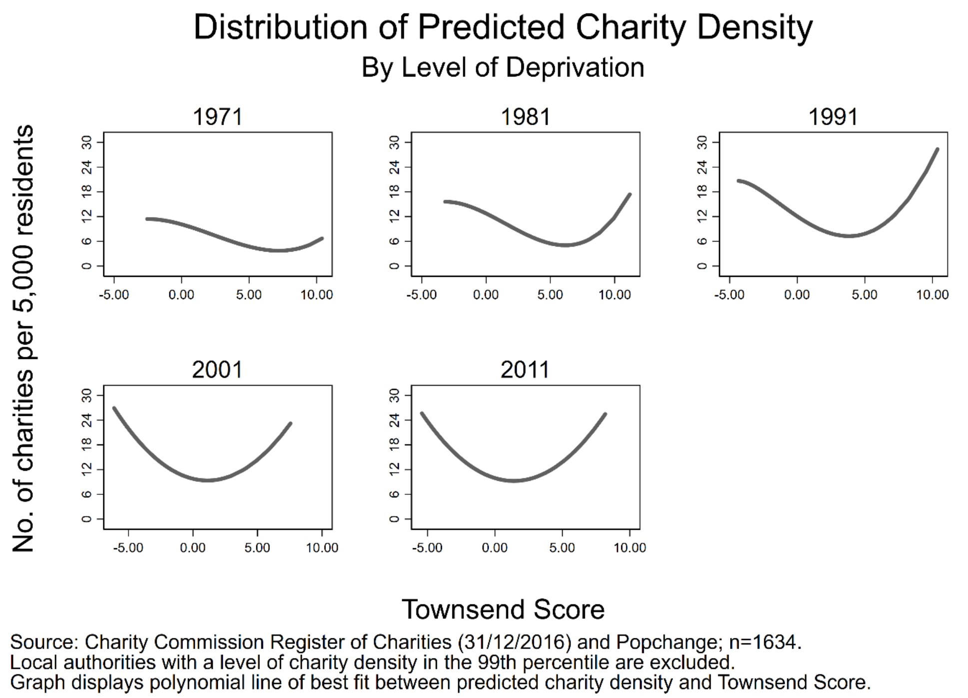


Figure 3.5 presents the temporal variation in the association between charity density and the level of deprivation in a local authority. Not only can we make comparisons between local authorities in a given year, we can now examine change *over time*, adding much more detail to our understanding of the relationship between density and deprivation.

Some research questions require longitudinal data.

Figure 1.6.

		Change in charity density			
Factors		1981	1991	2001	2011
Townsend Score ranking					
	Less deprived	REF	REF	REF	REF
	More deprived	-.34*	.28**	.39*	.43*
Urban/rural classification					
	More urban	-.78**	.72	-.30	-.72
	More rural	-.78**	.09	-.46	.73***
Previous charity density		-	.71***	0.04	-.41***
N (local authorities)		317	316	319	318
Adjusted R^2		.02	.56	.01	.20
F test		5.23***	41.43***	1.59	21.08***

Figure 3.3 displays the results of a change score model that links changes in the values of a set of explanatory variables to changes in the values of the outcome. For example, a local authority becoming more deprived between census years is associated with a small increase in charity density. Such an analysis is not possible if we did not have data on the same local authorities at multiple time periods.

Research questions that require longitudinal data:

- Flows into and out of poverty.
- The effects of family migration on the woman's subsequent employment activities.
- The impact of Covid-19 on long-term health outcomes of individuals.
- Evaluating policy, health, educational interventions.

Methodological benefits

Micro-level social processes

Repeated cross-sectional data can reveal macro-level trends and patterns of substantive interest but mask micro-level change. For example, repeated cross-sectional analysis of the British Household Panel Survey (precursor to Understanding Society) showed that poverty rates stabilised in the 1990s. However longitudinal analysis uncovered substantial turnover / churn in terms of which individuals remained in or exited poverty (the poor are not always poor!).

In my own research area, cross-sectional analysis of the Scottish Household Survey reveals the proportion of individuals volunteering has remained stable between 2007-2017 (Volunteer Scotland, 2019). However this pattern masks the substantial micro-level variation in volunteering behaviour: that is, it is not the same individuals volunteering every year, with people dipping in and out of this activity throughout the lifecourse.

Temporal ordering of events

Longitudinal data give us a better sense of the timing of events and hence the direction of influence. Remember that a necessary (but insufficient) condition for causal analysis is the appropriate temporal ordering of the cause and effect: X cannot cause Y if it does not occur before Y.

Understanding — and having the ability to identify — the temporal ordering of events helps to address a pervading issue in quantitative social science analysis: *simultaneity bias*. For example, it is difficult to untangle whether poor health causes unemployment, unemployment causes poor health (or both) without some form of longitudinal data.

Improving control for residual heterogeneity

Now we arrive at one of the major methodological appeals of longitudinal data: the ability to control for *residual heterogeneity*. As Gayle (2018) concisely states:

> The possibility of substantial variation between similar individuals due to unmeasured, and possibly immeasurable, variables is known as 'residual heterogeneity'.

You may have heard residual heterogeneity referred to as *omitted variable bias* or *unobserved hetereogeneity*. We'll spend much more time on this benefit in the next section.

Improving control for state dependence

Longitudinal data provide important information on the initial or current state an entity is in, and the trajectory of said entity across different or the same states over time. As Nobel Prize winner J.J. Heckman summarises:

> A frequently notes empirical regularity in the analysis of employment 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.

In essence, much of human behaviour is influenced by previous behaviour and outcomes. Think back to the example we showed from the Millennium Cohort Study: both boys and girls were most likely to remain at the same weight (whether normal or excess) at age 14 as they were at age 11.

A note of caution

Longitudinal data are not a panacea:

- For missing data
- For measurement error

- For lack of sample representativeness
- For poorly specified statistical models
- Etc

See the excellent summaries of the strengths and weaknesses of longitudinal data produced by CLOSER.

In summary

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

What does longitudinal data look like?

Let's get our hands dirty working with some real-world longitudinal data: strictly speaking I'll get my hands dirty as the data set we're using has some restrictions on sharing. We will explore a data set containing a representative sample of UK charities: a version of this data set is available through the UK Data Service: SN 853257

First, let's start with a simple, fabricated example of a longitudinal data set.

```
import delimited using "./data/lda-simple-example-2020-08-28.csv", clear varn(1)
l
```

```
(5 vars, 20 obs)
```

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

     |   pid   year      sex   age   income |

     |--------------------------------------|

  1. | 10001   2015     male    22    20000 |

  2. | 10001   2016     male    23    20000 |

  3. | 10001   2017     male    24    22000 |

  4. | 10001   2018     male    25    24000 |

  5. | 10002   2015   female    45    29000 |

     |--------------------------------------|

  6. | 10002   2016   female    46    29000 |

  7. | 10002   2017   female    47    29000 |

  8. | 10002   2018   female    48    29500 |

  9. | 10003   2015   female    31    41500 |

 10. | 10003   2016   female    32    42400 |

     |--------------------------------------|

 11. | 10003   2017   female    33    43800 |

 12. | 10003   2018   female    34    45000 |
```

```
 13. | 10004   2015     male    65    25000 |
```

```
 14. | 10004   2016     male    66    10000 |
```

```
 15. | 10004   2017     male    67    10000 |
```

```
     |-------------------------------------|
```

```
 16. | 10004   2018     male    68    10000 |
```

```
 17. | 10005   2015   female    18    14000 |
```

```
 18. | 10005   2016   female    19    15000 |
```

```
 19. | 10005   2017   female    20    15000 |
```

```
 20. | 10005   2018   female    21    18000 |
```

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

Here we have five individuals (*units*) observed across four years (*time periods*), with three variables capturing attributes in each year (`sex`, `age`, `income`).

This is an example of a **balanced panel**: the same number of observations is captured for each unit.

Now let's look at a different example:

```
import delimited using "./data/lda-simple-example-ub-2020-08-28.csv", clear varn(1)
l
```

```
(5 vars, 16 obs)
```

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

```
     |   pid   year      sex   age   income |
```

```
     |-------------------------------------|
```

```
  1. | 10001   2015     male    22    20000 |
```

```
  2. | 10001   2016     male    23    20000 |
```

```
  3. | 10001   2017     male    24    22000 |
```

```
  4. | 10001   2018     male    25    24000 |
```

```
  5. | 10002   2015   female    45    29000 |
```

```
     |-------------------------------------|
```

```
  6. | 10002   2016   female    46    29000 |
```

```
  7. | 10003   2015   female    31    41500 |
```

```
  8. | 10003   2016   female    32    42400 |
```

```
  9. | 10003   2017   female    33    43800 |

 10. | 10004   2015     male    65    25000 |

     |------------------------------------|

 11. | 10004   2016     male    66    10000 |

 12. | 10004   2017     male    67    10000 |

 13. | 10004   2018     male    68    10000 |

 14. | 10005   2015   female    18    14000 |

 15. | 10005   2016   female    19    15000 |

     |------------------------------------|

 16. | 10005   2017   female    20    15000 |

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

Here we have the same units and time span but this time there are gaps within units: individual `10002` is only observed twice, and `10003` and `10005` three times.

This is an example of an **unbalanced panel**: the same number of observations is not captured for each unit.

Working with a balanced panel is preferrable for a number of reasons, which we'll explore in due course. However the methods of analysis we will cover apply to unbalanced panels also (Mehmetoglu & Jakobsen, 2016).

The classic panel consists of a large number of units of analysis (*i*) observed over a small number of periods (*t*).

Charity data

```
use "./data/charity-panel-2020-09-10.dta", clear
desc
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)

Contains data from ./data/charity-panel-2020-09-10.dta

  obs:        68,818                          Contains annual accounts of

                                                charities in E&W for financial

                                                years 2006-2017

 vars:            31                          9 Sep 2020 08:41

 size:     8,326,978                          (_dta has notes)

--------------------------------------------------------------------------------

              storage   display    value

variable name   type    format     label      variable label
```

```
-------------------------------------------------------------------------------

regno           long    %12.0g                Charity number (unique id)

fin_year        byte    %8.0g      fin_year   Financial year

etotal          double  %12.0g                Total expenditure

itotal          double  %12.0g                Total income

aob_classified  str19   %19s                  Geographical scale of activity

                                                i.e. local, national

sampling_strata byte    %12.0g     sampling_strata_lab

                                              Income categories used to sample

                                                organisations

large_samplin~a byte    %12.0g     large_sampling_strata_lab

                                              Income categories used to sample

                                                large organisations (£500k+)

orgsize         byte    %12.0g     orgsize_lab

                                              Size of charity - in categories of

                                                total annual gross income

orgsize_large   byte    %12.0g     orgsize_large_lab

                                              Organisation size by income bands,

                                                for large charities (> £500k)

orgsize_alt     byte    %13.0g     orgsize_alt_lab

                                              Organisation size by income bands,

                                                alternative banding

fundraised      float   %9.0g                 Income derived from donations from

                                                individuals

ind_fees        float   %9.0g                 Income derived from fees for

                                                charitable activities from

                                                individuals
```

govern	float	%9.0g	Income derived from government
			grants or contracts
volsector	float	%9.0g	Income derived from voluntary
			sector grants or contracts
internal	float	%9.0g	Income derived from investments
			and trading subsidiaries
business_other	float	%9.0g	Income derived from other sources
			e.g. business sector
fundraised_sh~e	float	%9.0g	Share of income derived from
			donations from individuals
business_othe~e	float	%9.0g	Share of income derived from other
			sources e.g. business sector
internal_share	float	%9.0g	Share of income derived from
			investments and trading
			subsidiaries
volsector_share	float	%9.0g	Share of income derived from
			voluntary sector grants or
			contracts
govern_share	float	%9.0g	Share of income derived from
			government grants or contracts
ind_fees_share	float	%9.0g	Share of income derived from fees
			for charitable activities from
			individuals
nsources	byte	%9.0g	Number of income sources where
			income >= £1,000
inc_diverse	float	%9.0g	Index of revenue diversification:
			0 (less diversified) to 1 (more

```
                                              diversified)
```

```
maxyear         byte    %9.0g                Most recent year charity appears
```

```
                                               in the dataset
```

```
orgage          int     %9.0g                Age of charity - in years
```

```
linc            float   %9.0g                Total income (log)
```

```
genchar         float   %9.0g                General charity
```

```
socser          float   %9.0g                Social service charity
```

```
west            float   %9.0g                Charity registered in Westminster
```

```
localc          float   %9.0g                Local charity
```

```
--------------------------------------------------------------------------------
```

```
Sorted by: regno
```

Let's perform a couple of quick tasks in order to get familiar with the data.

First, we need to tell Stata we are dealing with panel data, as this allows us to access some time-series operators that are useful:

```
xtset regno fin_year
```

```
       panel variable:  regno (unbalanced)
```

```
        time variable:  fin_year, 1 to 11, but with gaps
```

```
                delta:  1 unit
```

The `xtset` command takes two arguments: a variable representing the unique identifier of the panel units (`regno`) and a variable capturing the unique identifier for the time period (`fin_year`). This combination of variables must uniquely identify every observation (row) in the data: we can check whether this is the case using the `isid` command - if no error message is returned, then those variables uniquely identify an every observation:

```
isid regno fin_year
```

Second, we can use `xtdescribe` to learn more about the patterns of observations in our panel:

```
xtdescribe
```

```
   regno:  200048, 200051, ..., 1166968                     n =      11193
```

```
fin_year:  1, 2, ..., 11                                    T =         11
```

```
           Delta(fin_year) = 1 unit
```

```
           Span(fin_year)  = 11 periods
```

```
           (regno*fin_year uniquely identifies each observation)
```

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

```
                           1       1       3         6         10       11       11

      Freq.  Percent    Cum. |  Pattern

  ---------------------------+-------------

      2166     19.35   19.35 |  11111111111

       476      4.25   23.60 |  ..111111111

       434      3.88   27.48 |  ....1.1.1.1

       388      3.47   30.95 |  ........1.1

       381      3.40   34.35 |  ......1.1.1

       247      2.21   36.56 |  ....1......

       212      1.89   38.45 |  .......1.1.

       211      1.89   40.34 |  ......1....

       181      1.62   41.95 |  1111.......

      6497     58.05  100.00 | (other patterns)

  ---------------------------+-------------

     11193    100.00         |  XXXXXXXXXXX
```

Let’s unpack these results:

- There are 11,193 panel units (*n*) and 11 time periods (*T*).
- The time period variable (*fin_year*) changes by 1 unit (*Delta(fin_year)*).
- 50% of panel units are observed at least 6 times (*Distribution of T_i*).
- 2,166 panel units are observed in every time period, 181 are observed only in the first 4 periods etc (see frequency table).

```
by regno: gen numobs = _N
xttab numobs
```

```
                    Overall             Between            Within

    numobs |     Freq.  Percent      Freq.  Percent        Percent

 ----------+-----------------------------------------------------

         1 |     1318      1.92      1318     11.78         100.00

         2 |     2838      4.12      1419     12.68         100.00

         3 |     3069      4.46      1023      9.14         100.00

         4 |     3812      5.54       953      8.51         100.00

         5 |     2895      4.21       579      5.17         100.00
```

```
         6 |      3624       5.27        604       5.40          100.00
         7 |      4081       5.93        583       5.21          100.00
         8 |      4552       6.61        569       5.08          100.00
         9 |      8883      12.91        987       8.82          100.00
        10 |      9920      14.41        992       8.86          100.00
        11 |     23826      34.62       2166      19.35          100.00
----------+----------------------------------------------------
     Total |     68818     100.00      11193     100.00          100.00
                              (n = 11193)
```

Now we have a better sense of the number of times we observe our panel units in the data. Let's also create a variable that identifies charities that appear in every year in the data, and drop all charities that do not meet this criterion:

```
gen balpan = (numobs==11)
keep if balpan
```

```
(44,992 observations deleted)
```

That will do for now, we'll examine the variables when we start estimating statistical models in the next section. We'll save the changes to the data set:

```
sav "./data/charity-panel-analysis-2020-09-10.dta", replace
```

```
file ./data/charity-panel-analysis-2020-09-10.dta saved
```

Summary

Longitudinal data offer a number of substantive and methodological benefits.

There a number of study designs, each with strengths and weaknesses.

Longitudinal data are not a panacea.

Panel Data Analysis I

In this section we define the general methodological and substantive issues associated with panel data.

We conclude with a consideration of the key questions a researcher should ask before undertaking analysis of panel data.

Introduction

The analysis of repeated contacts data is known as **panel data analysis**.

Recall that repeated contacts data captures information on your units of analysis more than once. As a result, observations are *nested* or *clustered* within units e.g., observations of pupils' exam results are nested within schools.

Methodological implications of panel data

The use of panel data implies the potential for the violation of an important regression assumption: error terms are independent of each other (Mehmetoglu & Jakobsen, 2016)

In panel data a unit's own observations are often *interdependent*, meaning they are more likely to be similar to each other than the observations for other units in the panel.

Independence of error term

Recall one of the core assumptions of linear regression:

$$\mathrm{cov}(\epsilon, X) = 0$$

The variation in our outcome that is left unexplained (ϵ) should not be correlated with any of the explanatory variables in the model.

If the covariance is **not equal** to zero, then the observations for each unit i are *serially correlated*, a circumstance also known as *autocorrelation*.

What this means in practice is the value of a variable at time t predicts the value of the same variable at time $t + k$ for a given unit i (where k represents another time period in which unit i is observed).

Autocorrelation can give rise to *heteroscedasticity*, which very often results in the under-estimation of standard errors in regression models.

It can also lead to the much more serious issue of biased coefficients.

Summary of issues

Panel data contain observations nested within units.

The interdependence of observations often violates a key assumption of linear regression (*independence of errors*).

Ignoring this interdependence when estimating your statistical model can lead to two problems:

1. Under-estimation of the uncertainty surrounding the coefficients (*inefficiency*).
2. Incorrect estimates of the coefficients (*bias*).

Inefficiency leads to under-estimated standard errors and potential false positive tests of statistical significance.

Bias leads to incorrect inferences about the magnitude and direction of the effects of the explanatory variables in your model.

Methodological benefits of panel data

Hold on, this entire training course is predicated on there being some advantage to using panel data over cross-sectional data!

Correct, and here it is…

The problem of **inefficient estimates** can at least be ameliorated when using cross-sectional data (e.g., robust or clustered standard errors).

The problem of **biased coefficients** is very difficult to solve when using cross-sectional data.

This because it is very difficult to find a data set that contains all of the explanatory variables you need for your model –> omitted variable bias.

Let's see what happens when omitted variable bias is present; that is, we have not specified the model correctly:

```
clear
capture set seed 1010
quietly set obs 10000

gen x1 = rnormal(1, 20)
gen x2 = x1 + rnormal(1, 10)
gen eterm = rnormal()
gen y = 2 + x1 + x2 + eterm
l y x1 x2 in 1/10
```

```
     +----------------------------------+
     |         y          x1          x2 |
     |----------------------------------|
  1. |  33.65662    19.14529    13.26858 |
  2. | -49.57088   -27.96305     -23.022 |
  3. |  13.81728     5.44816    4.905838 |
  4. | -18.24858   -4.415646    -16.3728 |
  5. |   25.3734    7.114079    16.31598 |
     |----------------------------------|
  6. |  41.18281     11.9115    26.35516 |
  7. | -45.91599   -18.31569   -28.86481 |
  8. | -17.55372   -6.058764   -11.95182 |
  9. |  47.78559    19.07098    27.25243 |
 10. |  11.26871    8.339461    1.703953 |
     +----------------------------------+
```

First, we estimate a properly specified model:

```
regress y x1 x2
```

```
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(2, 9997)      >  99999.00
       Model |  16796126.4         2  8398063.21   Prob > F        =    0.0000
    Residual |  10024.3942     9,997  1.00274025   R-squared       =    0.9994
-------------+----------------------------------   Adj R-squared   =    0.9994
       Total |  16806150.8     9,999  1680.78316   Root MSE        =    1.0014
------------------------------------------------------------------------------
```

```
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    1.00037   .0011332   882.78   0.000     .9981488    1.002591
          x2 |   .9994168   .0010176   982.18   0.000     .9974221    1.001411
       _cons |   1.989778   .0100956   197.09   0.000     1.969989    2.009568
------------------------------------------------------------------------------
```

Now let's estimate a model that excludes one of the explanatory variables:

```
regress y x1
```

```
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      >  99999.00
       Model |  15828807.8         1  15828807.8   Prob > F        =    0.0000
    Residual |  977343.026     9,998  97.7538533   R-squared       =    0.9418
-------------+----------------------------------   Adj R-squared   =    0.9418
       Total |  16806150.8     9,999  1680.78316   Root MSE        =    9.8871
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.997809   .0049647   402.40   0.000     1.988077    2.007541
       _cons |   3.076904   .0990786    31.06   0.000      2.88269    3.271118
------------------------------------------------------------------------------
```

Notice how the coefficient for `x1` has been inflated? This is because `x1` and `x2` are correlated (by definition), and therefore `x1` "soaks up" some of the variation in `y` that is explained by `x2` (Gelman and Hill, 2007).

```
corr x1 x2
corr y x2
```

```
(obs=10,000)
             |       x1       x2
-------------+------------------
          x1 |   1.0000
          x2 |   0.8962   1.0000
```

```
(obs=10,000)

             |        y       x2
-------------+------------------
           y |   1.0000
          x2 |   0.9762   1.0000

```

So why panel data?

As the simple example above demonstrates, one way of solving omitted variable bias is to include the omitted explanatory variable(s)!

This can be difficult to achieve in practice, as many of these variables may not be captured by the data set, or even possible to record at all (Mehmetoglu & Jakobsen, 2016).

If certain assumptions hold, the use of panel data allow us to control for the influence of any omitted variables on the coefficients of the explanatory variables.

Key assumption: the omitted variables are *time-invariant*.

> 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. (Gayle, 2018)

Panel data won't completely address this problem, but suitable models can improve control for, and even estimate the effects of, omitted explanatory variables.

Substantive benefits of panel data

It would be unwise to focus exclusively on the methodological implications of panel data.

A major advantage of such data sets is their ability to capture social processes as they evolve over time (*micro-level change*).

```
import delimited using "./data/lda-employed-example-2020-08-28.csv", clear varn(1)
tab pid employed
```

```
(3 vars, 20 obs)

           |       employed
       pid |         0          1 |     Total
-----------+----------------------+----------
     10001 |         5          5 |        10
     10025 |         5          5 |        10
-----------+----------------------+----------
     Total |        10         10 |        20
```

In this fictional example we see that the two individuals have the same overall employment history: five periods of employment, five of unemployment.

However this summary masks the stark difference in their employment trajectories:

```
1
```

```
     +------------------------+
     |   pid   year   employed |
     |------------------------|
  1. | 10001   2000          1 |
  2. | 10001   2001          1 |
  3. | 10001   2002          0 |
  4. | 10001   2003          1 |
  5. | 10001   2004          0 |
     |------------------------|
  6. | 10001   2005          1 |
  7. | 10001   2006          1 |
  8. | 10001   2007          0 |
  9. | 10001   2008          0 |
 10. | 10001   2009          0 |
     |------------------------|
 11. | 10025   2000          1 |
 12. | 10025   2001          1 |
 13. | 10025   2002          1 |
 14. | 10025   2003          1 |
 15. | 10025   2004          1 |
     |------------------------|
 16. | 10025   2005          0 |
 17. | 10025   2006          0 |
```

```
 18. | 10025   2007          0 |
```

```
 19. | 10025   2008          0 |
```

```
 20. | 10025   2009          0 |
```

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

Individual `10001` drifts in and out of employment, while `10025` only changes employment status once (in 2005).

Therefore we can decide to focus on analysing change over time, in addition to traditional analyses of differences between groups:

```
xtset pid year
bys pid: xttrans employed
```

```
       panel variable:  pid (strongly balanced)
```

```
        time variable:  year, 2000 to 2009
```

```
                delta:  1 unit
```

```
--------------------------------------------------------------------------------
```

```
-> pid = 10001
```

```
           |       employed
```

```
  employed |         0          1 |     Total
```

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

```
         0 |     50.00      50.00 |    100.00
```

```
         1 |     60.00      40.00 |    100.00
```

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

```
     Total |     55.56      44.44 |    100.00
```

```
--------------------------------------------------------------------------------
```

```
-> pid = 10025
```

```
           |       employed
```

```
  employed |         0          1 |     Total
```

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

```
         0 |    100.00       0.00 |    100.00
```

```
         1 |     20.00      80.00 |    100.00
```

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

```
     Total |     55.56      44.44 |    100.00
```

Panel data analysis: key considerations

How can we use our understanding of these two advantages of panel data — **examining micro-level change** and **improved control for residual heterogeneity** — when estimating statistical models?

A good approach is to pose two overarching questions:

How do your explanatory variables influence the outcome?

- Are you interested in how *changes within units* are associated with variation in the outcome?
- Are you interested in how *differences between units* are associated with variation in the outcome?
- Both?

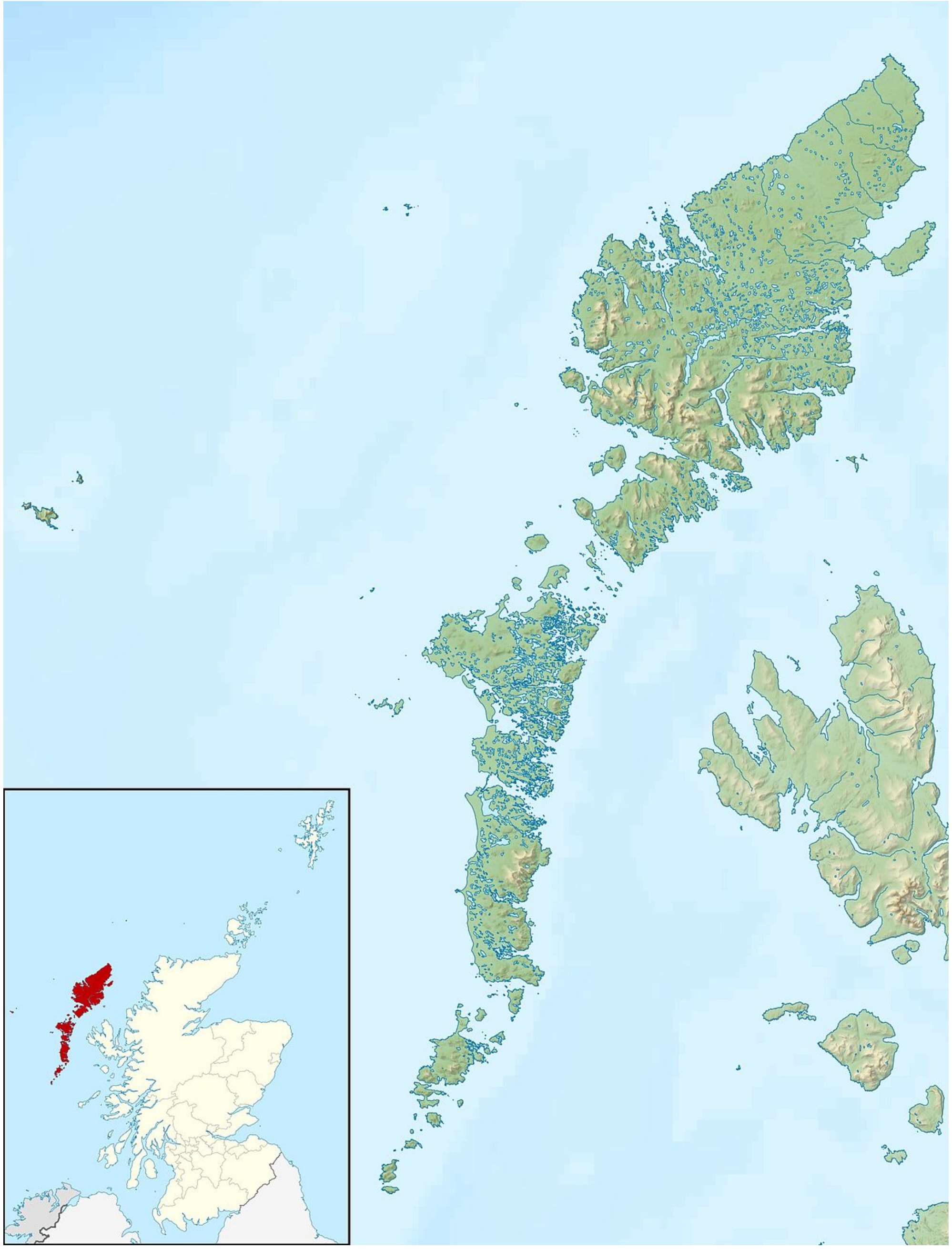
Consider this simple example:

Would you expect the effect of retirement on income to differ whether:

- we were comparing two individuals (one retired, one not), or
- we were comparing one individual who changes retirement status between two time periods?

Here is another example:

Average earnings in the Outer Hebrides of Scotland are lower than average for London. But would we expect earnings to drop on average if someone moves from London to the Outer Hebrides?



Credit: [Wikipedia]()

Answering the question — *how do your explanatory variables influence the outcome?* — requires theoretical insight on the nature of the relationships between your explanatory factors and outcome of interest. The decision you make influences which type of panel data model you ultimately select as being most appropriate for your research question.

Is your statistical model specified correctly?

Do you have all and only relevant explanatory variables in your model (Gelman and Hill, 2007)?

How worried are you that some (especially important) explanatory variables have not been included in your model?

Do you think the omission of these explanatory variables is leading to bias in the variables included in the model?

This is a technical issue and there are a number of statistical tests and techniques that can help guide us to select the most appropriate panel data model.

Task

Think of a piece of quantitative analysis you have done (or would like to do).

Clearly state the analysis in terms of an outcome and a set of explanatory variables (a statistical model).

Consider the two main questions:

- *How does each of your explanatory variables influence the outcome?*
- *Is your statistical model specified correctly?*

Finally, consider whether and how panel data would support the estimation of your statistical model.

Panel Data Analysis II

In this section we estimate our first set of statistical models using panel data: **Pooled OLS** and **Between Effects**. We show some examples of how to estimate and interpret these models, and reflect on the conditions under which the models are appropriate.

What we can relax about

In the sessions demonstrating how to quantitatively analyse panel data, we will cast aside the following concerns:

- Missing data
- Weights
- Attrition
- Multicollinearity

All of these issues impinge on the estimation of panel data models but are not necessary to address for the purposes of learning about said models. We encourage you to consult the [reading list] for suggestions of resources that cover these topics.

Defining our statistical model

Now we arrive at the interesting bit: estimating statistical models.

Let's return to our panel data on charities and define a statistical model for predicting a charity's annual gross income as a function of its age, the scale of its charitable activities, where it is located, what type of charity it is, and the number of sources of income it has, and the share of its income provided by government.

$$\mathrm{y}_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5it} + \beta_6 x_{6it} + \epsilon_{it} \tag{1.6}$$

Where:

y_{it} is log income for charity *i* at time *t*

β_0 is the constant term, which is our prediction for log income when the values of all other variables in the model are set to 0

x_{1it} captures the age of charity *i* at time *t*, and β_1 is the effect of this variable on the outcome

x_{2i} is a dummy variable identifying charities that operate at a local level

x_{3i} is a dummy variable identifying charities registered in Westminster

x_{4i} is a dummy variable identifying general charities

x_{5it} captures the number of sources of income for charity *i* at time *t*

x_{6it} captures the share its income charity *i* derives from government sources at time *t*

ϵ_{it} captures the residual for charity *i* at time *t* $(\mathrm{y}_{it} - \hat{y}_{it})$

Understanding sources of variation

Remember to keep in mind the two sources of variation that exist in panel data ([Gould, n.d.]):

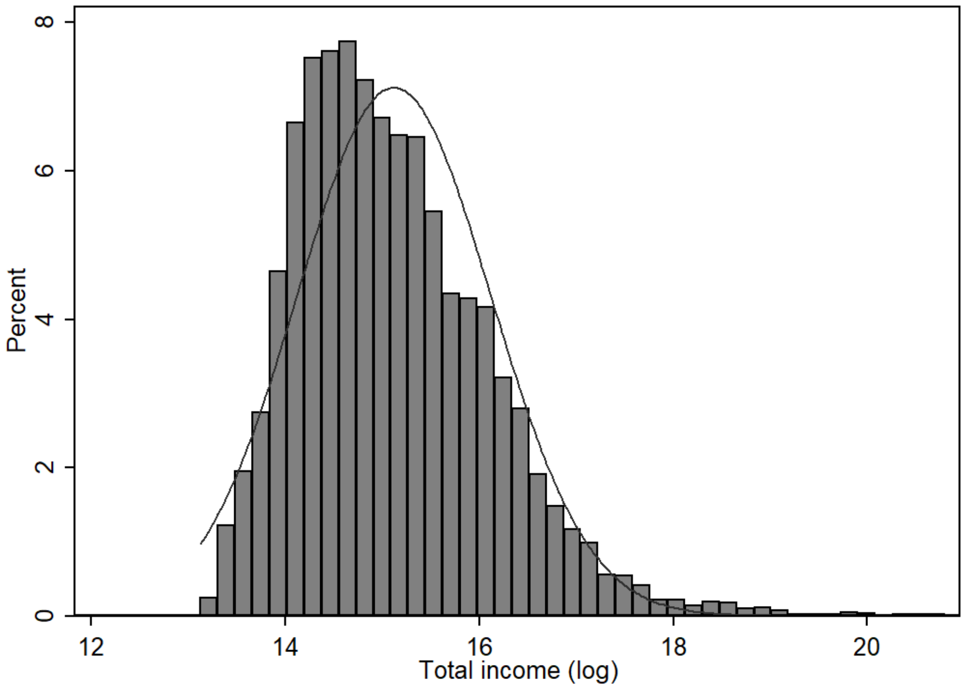
1. Cross-section information on differences between units
2. Time series information on differences over time within units

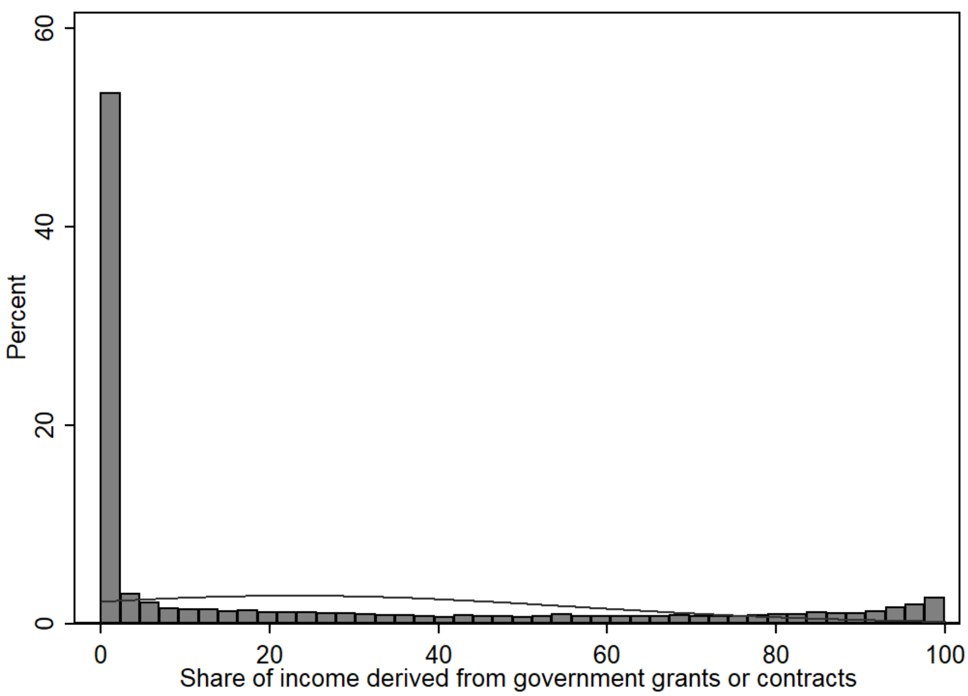
Data exploration

Let's spend a little bit of time exploring the key variables in our statistical model.

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)
```






```
sum orgage, detail
```

```
                    Age of charity - in years
```

```
-------------------------------------------------------------
      Percentiles      Smallest
 1%            4              0
 5%            7              1
10%           10              1       Obs              23,826
25%           16              1       Sum of Wgt.      23,826
50%           27                      Mean           39.20129
                        Largest       Std. Dev.       42.4661
75%           48            496
90%           82            497       Variance       1803.369
95%          112            498       Skewness       4.595531
99%          180            499       Kurtosis       37.17673
```

```
sum nsources, detail
```

```
      Number of income sources where income >= £1,000
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%            2              0
10%            2              1       Obs              23,826
25%            3              1       Sum of Wgt.      23,826
50%            4                      Mean           3.806724
                        Largest       Std. Dev.       1.24789
75%            5              6
90%            5              6       Variance       1.557228
95%            6              6       Skewness      -.1130695
99%            6              6       Kurtosis       2.425233
```

```
tab1 localc socser
```

```
-> tabulation of localc

      Local |

    charity |      Freq.     Percent        Cum.

------------+-----------------------------------

          0 |      8,756       36.75       36.75

          1 |     15,070       63.25      100.00

------------+-----------------------------------

      Total |     23,826      100.00

-> tabulation of socser

     Social |

    service |

    charity |      Freq.     Percent        Cum.

------------+-----------------------------------

          0 |     20,449       85.83       85.83

          1 |      3,377       14.17      100.00

------------+-----------------------------------

      Total |     23,826      100.00
```

Pooled OLS Model

The starting point for any statistical modelling of panel data is to estimate a *Pooled OLS* model (fancy way of saying linear regression).

The observations are "pooled", which just means we ignore the nested nature of panel data. In other words we assume that each observation (i.e., row within a long format data set) is independent of other observations (Gayle and Lambert, 2018).

Fundamental problem of pooling observations (Gayle & Lambert, 2018, p. 58):

> The model does not recognise that there are multiple contributions of data from the same individuals, and therefore, it estimates results as if there are many individuals who shared the same characteristics. This impacts upon the estimate of measures such as variances and standard errors.

```
regress linc orgage localc west genchar nsources govern_share
est store pols
```

```
      Source |       SS           df       MS      Number of obs   =    23,826

-------------+----------------------------------   F(6, 23819)     =    410.54
```

```
       Model |   2225.8864         6  370.981066   Prob > F        =    0.0000

    Residual |  21523.6961    23,819  .903635591   R-squared       =    0.0937

-------------+----------------------------------   Adj R-squared   =    0.0935

       Total |  23749.5825    23,825  .996834524   Root MSE        =     .9506

------------------------------------------------------------------------------

        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

-------------+----------------------------------------------------------------

      orgage |   .0036028     .00015    24.01   0.000     .0033087    .0038969

      localc |  -.3302434   .0130224   -25.36   0.000    -.3557682   -.3047187

        west |   .1121865   .0253139     4.43   0.000     .0625697    .1618033

     genchar |  -.3170303   .0139082   -22.79   0.000    -.3442913   -.2897693

    nsources |   .1053884   .0050963    20.68   0.000     .0953993    .1153774

govern_share |    .000644   .0002035     3.16   0.002     .0002451    .0010429

       _cons |   14.96317   .0236406   632.94   0.000     14.91683    15.00951

------------------------------------------------------------------------------

```

Conditions where Pooled OLS is suitable

Pooled OLS can produce consistent estimates of the explanatory variables if:

- The model is correctly specified
- The explanatory variables are uncorrelated with the error term (Cameron and Trivedi, 2010)

TASK: Do you think our statistical model is correctly specified, and there is no correlation between error term and explanatory variables?

In our statistical model of charity income, it is unlikely that the interpretation of the coefficients would change drastically if we addressed the under-estimation of the standard errors (the sample size is very large).

We'll cover the various tests and checks we can perform to examine whether Pooled OLS model violates the *independence of errors* assumption in a later section.

Between Effects Model

Once again estimate a cross-sectional model (Pooled OLS). However this time we transform the data so that there is one observation per unit. As a result we end up modelling the mean of Y on the mean of our X variables.

```
xtreg linc orgage localc west genchar nsources govern_share, be
est store beff
```

```
Between regression (regression on group means)  Number of obs     =     23,826
```

```
Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:

     within  = 0.0063                                         min =         11

     between = 0.1042                                         avg =       11.0

     overall = 0.0925                                         max =         11

                                                F(6,2159)         =      41.86

sd(u_i + avg(e_i.))=  .9109813                  Prob > F          =     0.0000

------------------------------------------------------------------------------

        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

-------------+----------------------------------------------------------------

      orgage |   .0035048   .0004791     7.32   0.000     .0025652    .0044443

      localc |  -.3282906   .0414364    -7.92   0.000      -.40955   -.2470312

        west |   .1167392   .0805284     1.45   0.147     -.041182    .2746604

     genchar |  -.3210918   .0449127    -7.15   0.000    -.4091685   -.233015

    nsources |   .1384596   .0193749     7.15   0.000     .1004643    .1764549

govern_share |   .0002749   .0007478     0.37   0.713    -.0011916    .0017415

       _cons |   14.85058   .0835091   177.83   0.000     14.68681    15.01435

------------------------------------------------------------------------------
```

Estimating a Between Effects model is equivalent to collapsing the data and estimating your regression model on the resulting observations:

```
preserve
    collapse (mean) linc orgage localc west genchar nsources govern_share, by(regno)
    regress linc orgage localc west genchar nsources govern_share
    est store coll
restore
```

```
      Source |       SS           df       MS      Number of obs   =     2,166

-------------+----------------------------------   F(6, 2159)      =     41.86

       Model |  208.427331         6  34.7378885   Prob > F        =    0.0000

    Residual |  1791.72593     2,159  .829886952   R-squared       =    0.1042

-------------+----------------------------------   Adj R-squared   =    0.1017
```

```
       Total |  2000.15326     2,165  .923858319   Root MSE        =    .91098
------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |   .0035048   .0004791     7.32   0.000     .0025652    .0044443
      localc |  -.3282906   .0414364    -7.92   0.000      -.40955   -.2470313
        west |   .1167393   .0805284     1.45   0.147    -.0411819    .2746605
     genchar |  -.3210918   .0449127    -7.15   0.000    -.4091685    -.233015
    nsources |   .1384596   .0193749     7.15   0.000     .1004643     .176455
govern_share |   .0002749   .0007478     0.37   0.713    -.0011916    .0017415
       _cons |   14.85058   .0835091   177.83   0.000     14.68681    15.01435
------------------------------------------------------------------------------

```

```
est table pols beff coll
```

```
----------------------------------------------------
    Variable |    pols         beff         coll
-------------+--------------------------------------
      orgage |  .00360282    .00350476    .00350476
      localc | -.33024344    -.3282906   -.32829062
        west |  .11218649    .11673923    .11673928
     genchar | -.31703032   -.32109178   -.32109176
    nsources |  .10538836     .1384596    .13845961
govern_share |  .00064402    .00027494    .00027494
       _cons |  14.963168    14.850581    14.850581
----------------------------------------------------
```

Benefits of Between Effects

- Sidesteps the problem of interdependence of observations in the original panel data.
- Smooths the effect of anomalous time periods (e.g., excess deaths calculation).
- Controls for omitted variables that change over time but are constant between units (e.g., national policies).

Limitations of Between Effects

What might the limitations of this approach be?

- Cannot estimate observed variables that change over time but are constant between units (e.g., national policies).
- Discard a lot of information by examining mean outcomes and inputs e.g., change over time.
- Cannot control for unobserved explanatory variables that are constant within but vary between units e.g., organisational culture.

The limitations of the Between Effects model far outweigh the benefits in most cases, and thus it is not widely used in practice (Mehmetoglu and Jakobsen, 2016). However it plays a crucial role in the estimation of another panel data model — Random Effects model — and thus it is important to understand how it works and what it offers.

Summary

Both the Pooled OLS and Between Effects models provide useful information on the association between an outcome *Y* and a set of explanatory variables *X*.

However both can be suboptimal from a substantive perspective (no change over time).

More concerningly, they offer no ability to control for residual heterogeneity in the form of *unobserved time-invariant* explanatory variables.

Panel Data Analysis III

In this section we estimate a statistical model that leverages some of the main advantages of using panel data: **Fixed Effects**. We show some examples of how to estimate and interpret this model, and reflect on the conditions under which the model is appropriate.

Quick reminder

Let's briefly recap some essential concepts regarding panel data:

Two sources of variation (Gould, n.d.):

1. Cross-section information on differences between units
2. Time series information on differences over time within units

So far our panel data models — Pooled OLS and Between Effects — only allow us to examine differences between units.

Two main issues with estimating statistical models:

1. Interdependence of errors
2. Improper model specification

The first can lead to inefficient estimates: under-estimated standard errors and false positive tests of statistical significance.

The second to biased coefficients and incorrect inferences regarding magnitude and direction of effect of explanatory variables.

Therefore we need a statistical model that allows us to **examine change over time** and/or **control for omitted variable bias**.

Defining our statistical model

Before estimating Fixed Effects and Random Effects models separately, it is worth identifying the key commonality between their respective statistical models.

Let's take a simplified version of our charity income statistical model, this time with only one explanatory variable (*age*) - typically it looks as follows:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \epsilon_{it} \tag{1.7}$$

However it is possible to **decompose** the residual variation (error term) into two separate terms:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \mu_i + e_{it} \tag{1.8}$$

In equation 1.8. we have introduced a **unit-specific** term to represent some of the residual variation in the outcome that is unexplained by the explanatory variables.

Decomposition implications

This term (μ_i) captures the effect of *residual heterogeneity* on the outcome i.e., unobserved or immeasurable characteristics of the units that are associated with the outcome (and possibly the explanatory variables), and vary across units.

In our charity data example, these charity-specific effects could be organisational culture, informal connections to government etc. In theory these characteristics could be measured but it's often wildly impractical.

The unit-specific effect also controls for the effect of other omitted variables on the outcome (and possibly the explanatory variables).

In our charity data example, we do not include explanatory variables capturing the amount a charity spends on fundraising, how well known it is etc.

A word of caution

Note the lack of a time subscript *t* in the new term μ_i. The implication is that the unobserved unit-specific effect is **constant** over time (i.e., within units).

Therefore Fixed Effects and Random Effects models only control for omitted variables that do not change within units (e.g., race, sex at birth, natural ability).

Fixed Effects Model

Conceptualising the Fixed Effects model

1. The Fixed Effects model focuses on how changes in explanatory variables are associated with changes in the outcome **within units**.
2. It assumes the observed explanatory variables and unobserved unit-specific effect are correlated (i.e., omitted variable bias is an issue).

Mehmetoglu and Jakobsen (2016, p. 241):

> "In other words, we use fixed effects whenever we are only interested in the impact of variables that vary over time. This estimator helps us explore the relationship between the dependent and the explanatory variables within a unit (person, company, country, etc.) Each unit has its own individual characteristics that may or may not influence the predictor variables."

The Fixed Effects model is specified as follows:

$$y_{it} = \beta_0 + \lambda_i + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + e_{it} \tag{1.9}$$

Where:

λ_i represents the unit-specific effect on the outcome.

The value of λ_i captures the effect of **all** of the unobserved time-invariant explanatory variables that are missing from the model. As a result, while the value of λ_i is calculated, it is not of much interest in and of itself. It's main role is to allow for a more robust (i.e., unbiased) estimation of the effects of the explanatory variables in the model.

In essence the Fixed Effects model produces a unit-specific intercept, which is the sum of the overall constant and the unit-specific effect:

$$y_{it} = \alpha_i + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + e_{it} \tag{1.10}$$

Where:

$\alpha_i = \beta_0 + \lambda_i$

The unit-specific effect shifts the overall intercept up or down the y axis by the value of λ.

Final thoughts on conceptualisation

Consider the Fixed Effects model a standard cross-sectional regression model with the addition of a dummy variable being for every unit in the panel except for one (i.e., *n - 1* dummy variables are added as explanatory variables).

Estimation

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)
```

```
xtreg linc orgage localc west genchar nsources govern_share, fe
```

```
note: localc omitted because of collinearity

note: west omitted because of collinearity

note: genchar omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =     23,826

Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:

     within  = 0.0140                                         min =         11

     between = 0.0425                                         avg =       11.0

     overall = 0.0403                                         max =         11

                                                F(3,21657)        =     102.28

corr(u_i, Xb)  = -0.1002                        Prob > F          =     0.0000

------------------------------------------------------------------------------

        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

-------------+----------------------------------------------------------------

      orgage |   .0069072   .0005802    11.90   0.000       .00577    .0080444

      localc |          0  (omitted)

        west |          0  (omitted)

     genchar |          0  (omitted)
```

```
    nsources |   .0289886   .0027931    10.38   0.000     .0235139    .0344633

govern_share |   .0010325   .0001225     8.43   0.000     .0007923    .0012727

       _cons |   14.71504    .026082   564.18   0.000     14.66392    14.76616

-------------+----------------------------------------------------------------

     sigma_u |  .94534636

     sigma_e |   .2821005

         rho |  .91823289   (fraction of variance due to u_i)

------------------------------------------------------------------------------

F test that all u_i=0: F(2165, 21657) = 120.80               Prob > F = 0.0000
```

QUESTION TIME

1. How much of the variation in the outcome is accounted for by the model? Is this a lot?
2. Why were three of the observed explanatory variables excluded in the estimation of the model?
3. What does the rho statistic tell us?
4. Is there evidence of correlation between the unit-specific effects and observed explanatory variables?

Interpretation

The effect of the observed explanatory variables is **net of** the effect of the unit-specific term. That is, we've controlled for the correlation between X and μ_i.

$\_\mathrm{cons}$ is the intercept and represents the average value of the fixed effects + the overall constant.

orgage is the predicted change in the outcome for a one-unit increase in organisational age.

rho is the proportion of unexplained variance in the outcome explained by unobserved differences between charities (the unit-specific effects), rather than changes within them.

If rho > .5 then most of the residual variation in the outcome is due to differences between units, if rho < .5 then most of the residual variation is accounted for by differences within units (i.e., the effects of the explanatory variables).

$\mathrm{corr(u\_i, Xb)}$ is the correlation between the unit-specific effect and the observed explanatory variables in the model.

- $\mathrm{sigma\_u}$ (or σ_u) is the standard deviation of the fixed effects (i.e., the residuals within units.
- $\mathrm{sigma\_e}$ (or σ_e) is the standard deviation of residuals ei.
- $\mathrm{R\text{-}sq{:}\ within}$ is the proportion of variance explained by the observed explanatory variables (i.e., excluding the unit-specific effect).

Post-estimation

Though it's very rarely of substantive interest, we can recover the unit-specific effects (and other parameter estimates) after estimating a Fixed Effects model:

```
capture predict fixed, u
capture predict y_hat, xb
capture predict ei, e
capture predict residuals, ue
capture egen pickone = tag(regno)
```

```
l regno fin_year fixed if pickone in 1/100
```

```
     +------------------------------+
     |  regno   fin_year      fixed |
     |------------------------------|
  1. | 200048    2006-07   -.9911593 |
 12. | 200051    2006-07    1.341765 |
 23. | 200069    2006-07   -.4608771 |
 34. | 200222    2006-07   -.1173254 |
 45. | 200424    2006-07   -.4077182 |
     |------------------------------|
 56. | 200431    2006-07   -.5248324 |
 67. | 200500    2006-07   -.0236679 |
 78. | 201081    2006-07    .0432656 |
 89. | 201321    2006-07   -1.400211 |
100. | 201911    2006-07   -.7076412 |
     +------------------------------+
```

```
l regno fin_year linc y_hat residuals fixed ei in 1/11
```

```
     +------------------------------------------------------------------+
  1. |  regno | fin_year |     linc |    y_hat | residuals |     fixed |
     | 200048 |  2006-07 | 14.00189 | 15.17772 | -1.175828 | -.9911593 |
     |------------------------------------------------------------------|
     |                                    ei                             |
     |                            -.1846687                              |
     +------------------------------------------------------------------+

     +------------------------------------------------------------------+
  2. |  regno | fin_year |     linc |    y_hat | residuals |     fixed |
     | 200048 |  2007-08 | 14.17788 | 15.18462 | -1.006747 | -.9911593 |
     |------------------------------------------------------------------|
```

```
     |                                    ei                            |
     |                             -.0155877                            |
     +------------------------------------------------------------------+
     +------------------------------------------------------------------+
  3. |  regno | fin_year |      linc |     y_hat | residuals |     fixed |
     | 200048 |  2008-09 |  14.1851 | 15.22075 |  -1.03565 | -.9911593 |
     |------------------------------------------------------------------|
     |                                    ei                            |
     |                             -.0444904                            |
     +------------------------------------------------------------------+
     +------------------------------------------------------------------+
  4. |  regno | fin_year |      linc |     y_hat | residuals |     fixed |
     | 200048 |  2009-10 |  14.2326 | 15.19844 | -.9658405 | -.9911593 |
     |------------------------------------------------------------------|
     |                                    ei                            |
     |                              .0253188                            |
     +------------------------------------------------------------------+
     +------------------------------------------------------------------+
  5. |  regno | fin_year |      linc |     y_hat | residuals |     fixed |
     | 200048 |  2010-11 |  14.1709 | 15.20695 | -1.036052 | -.9911593 |
     |------------------------------------------------------------------|
     |                                    ei                            |
     |                             -.0448926                            |
     +------------------------------------------------------------------+
     +------------------------------------------------------------------+
  6. |  regno | fin_year |      linc |     y_hat | residuals |     fixed |
     | 200048 |  2011-12 | 14.14801 | 15.21225 |  -1.06424 | -.9911593 |
```

```
     |---------------------------------------------------------------|
     |                                 ei                            |
     |                          -.0730808                            |
     +---------------------------------------------------------------+

     +---------------------------------------------------------------+
  7. |  regno | fin_year |     linc |    y_hat | residuals |     fixed |
     | 200048 |  2012-13 |   14.376 | 15.25097 | -.8749701 | -.9911593 |
     |---------------------------------------------------------------|
     |                                 ei                            |
     |                           .1161892                            |
     +---------------------------------------------------------------+

     +---------------------------------------------------------------+
  8. |  regno | fin_year |     linc |    y_hat | residuals |     fixed |
     | 200048 |  2013-14 | 14.29996 | 15.22607 | -.9261075 | -.9911593 |
     |---------------------------------------------------------------|
     |                                 ei                            |
     |                           .0650518                            |
     +---------------------------------------------------------------+

     +---------------------------------------------------------------+
  9. |  regno | fin_year |     linc |    y_hat | residuals |     fixed |
     | 200048 |  2014-15 | 14.26031 |  15.2623 | -1.001989 | -.9911593 |
     |---------------------------------------------------------------|
     |                                 ei                            |
     |                          -.0108296                            |
     +---------------------------------------------------------------+

     +---------------------------------------------------------------+
 10. |  regno | fin_year |     linc |    y_hat | residuals |     fixed |
```

```
     | 200048 |   2015-16 | 14.30113 | 15.23988 | -.9387508 | -.9911593 |
     |-----------------------------------------------------------------|
     |                                   ei                            |
     |                             .0524085                            |
     +-----------------------------------------------------------------+

     +-----------------------------------------------------------------+
 11. |  regno | fin_year |      linc |     y_hat | residuals |      fixed |
     | 200048 |   2016-17 | 14.37021 | 15.24679 | -.8765777 | -.9911593 |
     |-----------------------------------------------------------------|
     |                                   ei                            |
     |                             .1145816                            |
     +-----------------------------------------------------------------+
```

```
di -.9911593 + -.1846687
```

```
-1.175828
```

```
tabstat fixed ei, s(mean sd) format(%5.4f)
```

```
   stats |      fixed        ei
---------+--------------------
    mean |    -0.0000   -0.0000
      sd |     0.9451    0.2690
------------------------------
```

Benefits of Fixed Effects

- Analyse change over time.
- Control for residual heterogeneity.
- Coefficient estimates are **consistent** if the key assumption is true. That is, because we have controlled for the effect of unobserved time-invariant explanatory variables, our coefficients are more robust, which means increasing the sample size increases the likelihood the estimates are converging on their true values.

(Mehmetoglu and Jakobsen, 2016)

Limitations of Fixed Effects

- Ignores differences between units.
- Coefficient estimates are **inefficient**, especially when compared to those from a Random Effects model. As a result, standard errors tend to be larger. Put simply, the estimates of the coefficients are based on only one source of variation (within) and thus are more uncertain.

- Cannot include observed time-invariant explanatory variables. This is due to a very simple reason: if a value does not vary, how can it be associated with variation in the value of another variable?
- Cannot control for unobserved residual heterogeneity that varies over time e.g., educational ability? Natural resilience?
- It is not well suited for variables that rarely change within units.

Think carefully about variables that change little over time - how might these influence the outcome? For example, few individuals in your panel might switch from non-graduate to graduate (let's say you have a sample of older individuals). In a fixed effects model, your estimation of the effect of switching between non-graduate and graduate will be based on a small number of occurrences and care is due in interpreting the coefficient.

Summarising the Fixed Effects model

Focuses on change over time within a unit of analysis.

Can control for the effect of unobserved time-invariant explanatory variables (residual heterogeneity).

Provides robust estimates of observed explanatory variables when said variables are correlated with unobserved effects.

However cannot include observed explanatory variables that do not vary within units.

Summary

Both the Pooled OLS and Between Effects models provide useful information on the association between an outcome *Y* and a set of explanatory variables *X*.

Fixed Effects provide potentially different information on the association between an outcome *Y* and a set of explanatory variables *X.

Is there a way to combine the *within* and *between* perspectives?

Panel Data Analysis IV

In this section we estimate a statistical model that leverages some of the main advantages of using panel data: **Random Effects**. We show some examples of how to estimate and interpret this model, and reflect on the conditions under which the model is appropriate.

Quick reminder

Let's briefly recap some essential concepts regarding panel data:

Two sources of variation (Gould, n.d.):

1. Cross-section information on differences between units
2. Time series information on differences over time within units

Pooled OLS and Between Effects models only allow us to examine differences between units. Fixed Effects only allow us to examine differences within units.

What if we wanted a model that leveraged aspects of the Between Effects and Fixed Effects estimators? Such a model would give us:

- More variation with which to explain the outcome.
- More flexibilty (e.g., include observed time-invariant explanatory variables).
- Other methodological and modelling benefits (e.g., decomposing explanatory variables into within and between effects).

Defining our statistical model

Recall the general form of Fixed Effects and Random Effects models:

$$\mathrm{y}_{it} = \beta_0 + \beta_1 x_{1it} + \mu_i + \mathrm{e}_{it} \tag{1.8}$$

In equation 1.8. we have introduced a **unit-specific** term to represent some of the residual variation in the outcome that is unexplained by the explanatory variables.

This term (μ_i) captures the effect of *residual heterogeneity* on the outcome i.e., unobserved or immeasurable characteristics of the units that are associated with the outcome (and possibly the explanatory variables), and vary across units.

A word of caution

Note the lack of a time subscript *t* in the new term μ_i. The implication is that the unobserved unit-specific effect is **constant** over time (i.e., within units).

Therefore Fixed Effects and Random Effects models only control for omitted variables that do not change within units (e.g., race, sex at birth, natural ability).

Random Effects Model

Conceptualising the Random Effects model

1. The Random Effects model focuses on how changes in explanatory variables are associated with changes in the outcome **within and between units**.
2. It assumes the observed explanatory variables and unobserved unit-specific effects are **not** correlated. In essence: the unobserved unit-specific effect explains some of the variation in the outcome, but is not associated with any of the observed explanatory variables.

Gayle and Lambert (2018, p. 68):

> …the random effects panel model is a matrix weighted average of the within-effects (fixed effects) and the between effects.

In essence the Random Effects model borrows some information from the Between Effects model and some from the Fixed Effects model.

Therefore the coefficients in a Random Effects model allow you to speak in terms of the effect of an explanatory variable on an outcome, whether we are comparing different individuals or different observations for the same individual - we'll see what this means when we estimate our first Random Effects model.

The Random Effects model is specified as follows:

$$\mathrm{y}_{it} = \beta_0 + \beta_1 x_{1it} + \ldots + \beta_k x_{kit} + \upsilon_i + \mathrm{e}_{it} \tag{1.11}$$

Where:

υ_i represents the unit-specific effect on the outcome.

As we assume the observed explanatory variables and unobserved unit-specific effects are **not** correlated, there is no need to estimate υ_i as if it were an explanatory variable and hence why it is part of the error component of the model.

Remember, it would only need to be estimated in the model if it would alter the coefficients for the observed explanatory variables: we assume it wouldn't it. Therefore instead of estimating the value of υ_i using the data (as we would for the observed explanatory variables), we assume the unit-specific effects are drawn from a known probability distribution (Gayle and Lambert, 2018).

As a result, we are only interested in the variance of υ_i and the extent to which it accounts for variation in the outcome.

Final thoughts on conceptualisation

In a Random Effects model we are unconcerned with estimating the coefficient of the unit-specific effect. We simply want to know to what degree variation in these unit-specific effects is associated with variation in the outcome.

Estimation

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)
```

```
xtreg linc orgage localc west genchar nsources govern_share, re
```

```
Random-effects GLS regression                   Number of obs     =     23,826
Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:
     within  = 0.0135                                         min =         11
     between = 0.0888                                         avg =       11.0
     overall = 0.0832                                         max =         11

                                                Wald chi2(6)      =     507.12
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |    .005022   .0003686    13.62   0.000     .0042995    .0057446
      localc |  -.3323159   .0412995    -8.05   0.000    -.4132615   -.2513704
        west |   .0797179   .0801801     0.99   0.320    -.0774323    .2368681
     genchar |  -.2729578   .0414011    -6.59   0.000    -.3541025   -.1918131
    nsources |    .030451     .00276    11.03   0.000     .0250415    .0358605
govern_share |    .001024    .000121     8.47   0.000      .000787    .0012611
       _cons |   15.15988    .048456   312.86   0.000     15.06491    15.25485
-------------+----------------------------------------------------------------
     sigma_u |  .90700183
     sigma_e |   .2821005
         rho |  .91179586   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

QUESTION TIME

1. How much of the variation in the outcome is accounted for by the model? Is this a lot?
2. Why are `localc`, `west` and `genchar` included in the estimation of the model (when they weren't in the Fixed Effects model)?
3. What does the rho statistic tell us?
4. Is there evidence of correlation between the unit-specific effects and observed explanatory variables?

Interpretation

The effect of the observed explanatory variables is **not** net of the unit-specific effects. That is, we haven't controlled away any correlation between X and μ_i (because we assume they are not correlated).

_cons is the intercept.

orgage is the predicted change in the outcome for a one-unit increase in organisational age, **whether age changes within or between units**.

rho is the proportion of unexplained variance in the outcome explained by unobserved differences between charities (the unit-specific effects), rather than changes within them.

If rho > .5 then most of the residual variation in the outcome is due to differences between units, if rho < .5 then most of the residual variation is accounted for by differences within units (i.e., the effects of the explanatory variables).

$\text{corr(u\_i, X)} = 0$ is the formal statement of the (untested) assumption that there is no correlation between the unit-specific effect and the observed explanatory variables in the model.

- sigma_u (or σ_u) is the standard deviation of the fixed effects (i.e., the residuals within units.
- sigma_e (or σ_e) is the standard deviation of residuals ei.
- R-sq: overall is the proportion of variance explained by the observed explanatory variables (i.e., excluding the unit-specific effect).

Post-estimation

Though it's very rarely of substantive interest, we can recover the unit-specific effects (and other parameter estimates) after estimating a Random Effects model:

```
capture predict random, u
capture predict y_hat, xb
capture predict ei, e
capture predict residuals, ue
capture egen pickone = tag(regno)
```

```
l regno fin_year random if pickone in 1/100, clean
```

	regno	fin_year	random
1.	200048	2006-07	-1.144261
12.	200051	2006-07	1.582299
23.	200069	2006-07	-.21995
34.	200222	2006-07	-.199622
45.	200424	2006-07	-.1241415
56.	200431	2006-07	-.2140733
67.	200500	2006-07	.2228235
78.	201081	2006-07	.2918403

```
 89.   201321    2006-07    -1.13723
```

```
100.   201911    2006-07   -.6447842
```

```
l regno fin_year y_hat residuals random ei in 1/11, clean
```

```
        regno   fin_year      y_hat   residuals      random           ei
  1.   200048    2006-07   15.34991   -1.348024   -1.144261   -.2037623
  2.   200048    2007-08   15.35493   -1.177057   -1.144261   -.0327961
  3.   200048    2008-09   15.39064   -1.205535   -1.144261   -.0612741
  4.   200048    2009-10   15.36498   -1.132381   -1.144261    .0118806
  5.   200048    2010-11    15.3716   -1.200694   -1.144261   -.0564324
  6.   200048    2011-12   15.37502    -1.22701   -1.144261   -.0827486
  7.   200048    2012-13   15.41329   -1.037294   -1.144261    .1069672
  8.   200048    2013-14   15.38507   -1.085107   -1.144261    .0591543
  9.   200048    2014-15   15.42087   -1.160563   -1.144261   -.0163015
 10.   200048    2015-16   15.39511    -1.09398   -1.144261    .0502813
 11.   200048    2016-17   15.40013   -1.029922   -1.144261    .1143396
```

```
di -1.144261 + -.2037623
```

```
-1.3480233
```

```
tabstat random ei, s(mean sd) format(%5.4f)
```

```
   stats |    random        ei
---------+--------------------
    mean |    0.0000    0.0000
      sd |    0.9096    0.2691
------------------------------
```

Benefits of Random Effects

- Analyse both change over time and differences between units.
- Control for residual heterogeneity.
- Estimate observed time-invariant explanatory variables in the model.
- Coefficient estimates are **efficient**, especially when compared to those from a Fixed Effects model. As a result, standard errors tend to be smaller. Put simply, the estimates of the coefficients are based on more information than those in the Fixed Effects model, which bases its estimates on only one source of variation (within).

Limitations of Random Effects

- Key assumption is often unrealistic.
- Coefficient estimates are **inconsistent** if the key assumption is violated.
- That is, if the coefficients for the observed explanatory variables are biased, then increasing the sample size does not necessarily mean we are getting closer to the true value of the parameter. Difficult to infer whether the value of a coefficient is mainly determined by within or between variation (though there are solutions to this problem).
- Cannot control for unobserved residual heterogeneity that varies over time e.g., educational ability? Natural resilience?

The last point is worth expanding on: if units differ in an unobserved way that varies over time, this will not be controlled for in the Random Effects model.

Summarising the Random Effects model

Analyses both change within a unit's outcomes, and differences between units' outcomes.

Can control for the effect of unobserved time-invariant explanatory variables (residual heterogeneity).

Can include observed explanatory variables that do not vary within units (e.g., race, sex at birth).

Does not provide robust estimates of observed explanatory variables when said variables are correlated with unobserved unit-specific effects.

Summary

Both the Pooled OLS and Between Effects models provide useful information on the association between an outcome *Y* and a set of explanatory variables *X*.

Fixed Effects provide potentially different information on the association between an outcome *Y* and a set of explanatory variables *X.

Random Effects combines the *within* and *between* perspectives - methodological and substantive benefits.

Panel Data Analysis V

In this section we estimate a panoply of panel data models and try to determine which one is most appropriate for our data. We outline some tests — statistical and conceptual — that can be used to select from a set of panel data models.

Quick reminder

Let's quickly remind ourselves of the key questions we need to ask before estimating panel data models:

1. How do your explanatory variables influence the outcome?
2. Is your statistical model specified correctly?

Let's see how these questions map to the various panel data models we can estimate, and what tests we can run to help us select the most appropriate model (if it exists).

Defining our statistical model

Let's return to our panel data on charities and define a statistical model for predicting a charity's annual gross income as a function of its age, the scale of its charitable activities, where it is located, what type of charity it is, and the number of sources of income it has, and the share of its income provided by government.

$$\mathrm{y}_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5it} + \beta_6 x_{6it} + \epsilon_{it} \tag{1.6}$$

How do your explanatory variables influence the outcome?

We are interested in a model that allows us to include observed time-invariant explanatory variables, as these are of substantive interest. For example, are local charities typically smaller than national or international organisations?

It is possible that the effect of the observed time-varying explanatory variables may differ depending on whether we consider them from a within or between perspective. For example, the effect of gaining an additional income source — say new funding from government — may be different for a change within an individual charity than a comparison of two charities.

Is your statistical model specified correctly?

We would be surprised if there wasn't a correlation between the observed explanatory variables and the unobserved unit-specific effects. We only have six observed explanatory variables in the model, of which some do not vary much within charities (e.g., number of income sources), and some do not vary much between charities (e.g., a charity is either a social services organisation or not).

So before estimating models, we clearly want one that includes **observed time-invariant explanatory variables** and addresses the likely violation of **independence of errors** assumption.

Estimating models

Pooled OLS

Is the Pooled OLS model appropriate? That is, can we ignore the fact that charities likely differ in unobserved ways?

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)
```

```
regress linc orgage localc west genchar nsources govern_share
est store pols
```

```
      Source |       SS           df       MS      Number of obs   =    23,826
-------------+----------------------------------   F(6, 23819)     =    410.54
       Model |   2225.8864         6  370.981066   Prob > F        =    0.0000
    Residual |  21523.6961    23,819  .903635591   R-squared       =    0.0937
-------------+----------------------------------   Adj R-squared   =    0.0935
       Total |  23749.5825    23,825  .996834524   Root MSE        =     .9506
------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |   .0036028     .00015    24.01   0.000     .0033087    .0038969
      localc |  -.3302434   .0130224   -25.36   0.000    -.3557682   -.3047187
        west |   .1121865   .0253139     4.43   0.000     .0625697    .1618033
     genchar |  -.3170303   .0139082   -22.79   0.000    -.3442913   -.2897693
```

```
    nsources |   .1053884   .0050963    20.68   0.000     .0953993    .1153774
```

```
govern_share |    .000644   .0002035     3.16   0.002     .0002451    .0010429
```

```
       _cons |   14.96317   .0236406   632.94   0.000     14.91683    15.00951
```

```
------------------------------------------------------------------------------
```

We can perform an *autocorrelation* test to check whether *independence of errors* assumption is violated:

```
*net sj 3-2 st0039
*net install st0039

xtserial linc orgage localc west genchar nsources govern_share
```

```
Wooldridge test for autocorrelation in panel data
```

```
H0: no first-order autocorrelation
```

```
    F(  1,    2165) =    114.998
```

```
           Prob > F =      0.0000
```

The results of the Wooldridge strongly suggest the error terms are correlated across observations. In practice this means that the values of these variables typically vary less *within* than across units. An obvious example would be the `orgage` variable:

```
l regno orgage in 1/11, clean
```

```
        regno   orgage
```

```
  1.   200048       46
```

```
  2.   200048       47
```

```
  3.   200048       48
```

```
  4.   200048       49
```

```
  5.   200048       50
```

```
  6.   200048       51
```

```
  7.   200048       52
```

```
  8.   200048       53
```

```
  9.   200048       54
```

```
 10.   200048       55
```

```
 11.   200048       56
```

```
tabstat orgage, s(min max)
```

```
    variable |       min       max
-------------+--------------------
      orgage |         0       499
----------------------------------
```

```
xtsum orgage
```

```
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
orgage   overall |  39.20129    42.4661          0        499 |     N =   23826
         between |              42.35708         5        494 |     n =    2166
         within  |              3.162344  34.20129   44.20129 |     T =      11
```

- Overall results suggest the average age of a charity 39.
- Between results collapse data set down to one row per unit, hence slightly different figures to overall results. Min and Max now refer to average values.
- Within results calculate differences between observed value for a unit in a given period and the unit's mean value across all periods (and the global mean also, hence why results are counter-intuitive).

The presence of serial (auto) correlation suggests we cannot ignore the panel component of the data. However, that does not mean we need to estimate a panel model. We could use the `regress, cluster()` approach to relax the assumption that the error terms are independent and uncorrelated with the explanatory variables.

```
regress linc orgage localc west genchar nsources govern_share, cluster(regno)
```

```
Linear regression                               Number of obs     =     23,826
                                                F(6, 2165)        =      39.07
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0937
                                                Root MSE          =      .9506

                                 (Std. Err. adjusted for 2,166 clusters in regno)
------------------------------------------------------------------------------
             |               Robust
        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |   .0036028   .0005828     6.18   0.000     .0024599    .0047458
      localc |  -.3302434   .0451017    -7.32   0.000    -.4186907   -.2417962
```

```
        west |   .1121865   .0896089     1.25   0.211    -.0635419    .2879149
     genchar |  -.3170303   .0455036    -6.97   0.000    -.4062657   -.2277949
    nsources |   .1053884    .013709     7.69   0.000     .0785042    .1322725
govern_share |    .000644   .0006234     1.03   0.302    -.0005785    .0018665
       _cons |   14.96317   .0686493   217.97   0.000     14.82854    15.09779
------------------------------------------------------------------------------
```

We no longer have underestimated standard errors, resulting in more more accurate t tests of the coefficients (note some variables are no longer statistically significant). However we may still want to control for unit-specific differences in the outcome — that is, is some of the variation in the outcome explained by unobserved heterogeneity?

We can check whether a Random Effects model is preferred over Pooled OLS by conducting a *Breusch and Pagan Lagrangian multiplier test*.

```
quietly xtreg linc orgage localc west genchar nsources govern_share, re
xttest0
```

```
Breusch and Pagan Lagrangian multiplier test for random effects

        linc[regno,t] = Xb + u[regno] + e[regno,t]

        Estimated results:

                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                    linc |   .9968345        .998416
                       e |   .0795807       .2821005
                       u |   .8226523       .9070018

        Test:   Var(u) = 0

                             chibar2(01) = 98352.33
                          Prob > chibar2 =   0.0000
```

Rejection of the null hypothesis suggests that there is a panel effect on the outcome, and that a Random Effects model is preferred over Pooled OLS.

Fixed Effects or Random Effects?

For most repeated contacts data sets, it would be erroneous to ignore the panel component of the data, even after controlling for autocorrelation of the error terms.

We then have a choice between Fixed Effects and Random Effects. (We ignore the Between Effects model as it is rarely insightful on its own, and is captured by the Random Effects model anyway.)

Hausman Test

The *Hausman test* checks whether the coefficients of the Random Effects model are consistent — that is, equivalent to those from the Fixed Effects model.

Failure to reject the null hypothesis (they are equivalent) provides evidence in favour of the Random Effects model, otherwise the Fixed Effects model is considered more appropriate.

```
quietly xtreg linc orgage localc west genchar nsources govern_share, fe
est store fixed

quietly xtreg linc orgage localc west genchar nsources govern_share, re
est store random

hausman fixed random
```

```
                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     fixed        random       Difference          S.E.
-------------+----------------------------------------------------------------
      orgage |    .0069072      .005022        .0018852         .000448
    nsources |    .0289886      .030451       -.0014624        .0004286
govern_share |    .0010325      .001024        8.44e-06        .0000196
------------------------------------------------------------------------------
                           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)
                          =       53.31
                Prob>chi2 =      0.0000
```

In our example, it appears that the coefficients from the Random Effects model are inconsistent and thus the Fixed Effects model should be preferred.

Often you'll find that the *Hausman test* favours the Fixed Effects model but this isn't definitive proof that it is more appropriate.

Guidance on selecting an appropriate model

Confusing and conflicting advice is found throughout the statistical literature (Gelman and Hill, 2007).

In quantitative social science there is probably more support for Random Effects lately. Clark et al. (2010) state that Fixed Effects has its advantages but it limits the type of research questions that can be addressed.

Random Effects has qualities close to Fixed Effects where rich data are available i.e., where lots of observed time-varying explanatory variables are captured (Gayle and Lambert, 2018).

Selecting a model should first-and-foremost draw on theoretical insight on the relationship between the explanatory variables and the outcome.

Undertake the *Hausman test* but don't be bound by it (Gayle and Lambert, 2018).

Estimate theoretically plausible statistical models and carefully compare their results.

Summary

QUESTION

Which model of charity income would you choose and why?

Based on all of the guidance and the results of the statistical tests, I selected the Random Effects model.

Extensions

In this section we provide a whistle-stop tour of some additional techniques and approaches for panel data and longitudinal data more broadly.

Nonlinear outcomes

Fixed Effects and Random Effects models can be applied to nonlinear outcomes (e.g., binary and count dependent variables) also.

Here is a published example from McDonnell (2017): https://doi.org/10.1177/0899764017692039

```
use "./data/improvingcharityaccountability_20170411.dta", clear

gen localc = (geographicalspread==2)
gen linc = ln(totalfunds) if totalfunds > 0 & totalfunds!=.
```

```
(Scottish Charity Financial Exceptions Data: 2007-2013)


(1,323 missing values generated)
```

```
tab yearsubmitted excgroup_3
```

```
      Year |  Possible failure to
    annual |    apply funds for
    return |  charitable purposes
 submitted |         0          1 |     Total
-----------+----------------------+----------
      2007 |       754        196 |       950
      2008 |     2,752        881 |     3,633
      2009 |     2,964        818 |     3,782
      2010 |     2,946        736 |     3,682
      2011 |     2,659        702 |     3,361
      2012 |     2,450        645 |     3,095
      2013 |     1,555        457 |     2,012
      2014 |       585        222 |       807
-----------+----------------------+----------
     Total |    16,665      4,657 |    21,322
```

```
xtlogit excgroup_3 concentration charityage localc linc, or re
```

```
Fitting comparison model:

Iteration 0:   log likelihood = -10687.029
Iteration 1:   log likelihood = -10488.084
Iteration 2:   log likelihood = -10486.442
Iteration 3:   log likelihood = -10486.442

Fitting full model:

tau =  0.0     log likelihood = -10486.442
tau =  0.1     log likelihood = -10257.949
tau =  0.2     log likelihood =  -10048.82
tau =  0.3     log likelihood = -9859.3094
tau =  0.4     log likelihood =  -9689.005
tau =  0.5     log likelihood = -9538.3489
tau =  0.6     log likelihood = -9410.3167
tau =  0.7     log likelihood = -9314.0716
tau =  0.8     log likelihood =  -9277.005

Iteration 0:   log likelihood = -9313.6924
Iteration 1:   log likelihood = -9178.5858
Iteration 2:   log likelihood = -9173.5625
Iteration 3:   log likelihood = -9173.5365
Iteration 4:   log likelihood = -9173.5365  (backed up)
Iteration 5:   log likelihood = -9173.5362

Random-effects logistic regression              Number of obs     =     19,982
Group variable: org_id                          Number of groups  =      4,714

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =          1
                                                              avg =        4.2
                                                              max =          7

Integration method: mvaghermite                 Integration pts.  =         12

                                                Wald chi2(4)      =     232.05
Log likelihood  = -9173.5362                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
   excgroup_3 |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
concentration |   .8526391    .128419    -1.06   0.290     .6346916    1.145428
   charityage |   .9911255   .0015465    -5.71   0.000     .9880991    .9941612
       localc |    2.33282   .2335039     8.46   0.000     1.917256    2.838458
         linc |   1.333225   .0276658    13.86   0.000     1.280089    1.388567
        _cons |   .0050338   .0013577   -19.62   0.000      .002967    .0085406
-------------+----------------------------------------------------------------
     /lnsig2u |   1.518384   .0552301                      1.410135    1.626633
-------------+----------------------------------------------------------------
      sigma_u |   2.136549   .0590009                      2.023984    2.255376
          rho |   .5811599   .0134437                      .5546033    .6072544
------------------------------------------------------------------------------
LR test of rho=0: chibar2(01) = 2625.81                Prob >= chibar2 = 0.000
```

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear

xtpoisson nsources linc orgage localc west genchar govern_share, re
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)


Fitting Poisson model:

Iteration 0:   log likelihood =  -42473.77
Iteration 1:   log likelihood =  -42473.77

Fitting full model:

Iteration 0:   log likelihood = -43378.386
Iteration 1:   log likelihood = -41912.848  (not concave)
Iteration 2:   log likelihood = -41494.954
Iteration 3:   log likelihood = -41471.918
Iteration 4:   log likelihood = -41471.687
Iteration 5:   log likelihood = -41471.687

Random-effects Poisson regression               Number of obs     =     23,826
Group variable: regno                           Number of groups  =      2,166

Random effects u_i ~ Gamma                      Obs per group:
                                                              min =         11
                                                              avg =       11.0
                                                              max =         11

                                                Wald chi2(6)      =     231.78
Log likelihood  = -41471.687                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
    nsources |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        linc |   .0439247   .0057484     7.64   0.000     .0326581    .0551913
      orgage |   .0002729    .000149     1.83   0.067    -.0000192    .0005649
      localc |   .0063181   .0127513     0.50   0.620    -.0186741    .0313103
        west |  -.0484527   .0247255    -1.96   0.050    -.0969138    8.46e-06
     genchar |   .0671619   .0134216     5.00   0.000     .0408561    .0934677
govern_share |   .0016258   .0001569    10.36   0.000     .0013183    .0019333
       _cons |   .5776679   .0894589     6.46   0.000     .4023317    .7530042
-------------+----------------------------------------------------------------
    /lnalpha |  -2.928772   .0456227                      -3.01819   -2.839353
-------------+----------------------------------------------------------------
       alpha |   .0534627   .0024391                      .0488896    .0584635
------------------------------------------------------------------------------
LR test of alpha=0: chibar2(01) = 2004.16              Prob >= chibar2 = 0.000
```

Hybrid panel data models

A hybrid panel model allows you to decompose the observed explanatory variables into their within and between effects using the Random Effects estimator.

Let's return to our charity data example and see if we can decompose the effect of `nsources` into its within and between effects.

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)
```

```
bys regno: egen nsources_mn = mean(nsources)
gen nsources_delta = nsources - nsources_mn
```

```
xtreg linc orgage localc west genchar nsources_mn nsources_delta govern_share, re
```

```
Random-effects GLS regression                   Number of obs     =     23,826
Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:
     within  = 0.0136                                         min =         11
     between = 0.1017                                         avg =       11.0
     overall = 0.0952                                         max =         11

                                                Wald chi2(7)      =     536.49
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |   .0048981   .0003692    13.27   0.000     .0041745    .0056216
      localc |  -.3320839   .0412748    -8.05   0.000     -.412981   -.2511868
        west |   .1011212   .0802314     1.26   0.208    -.0561294    .2583718
     genchar |  -.3070555   .0418617    -7.34   0.000    -.3891029   -.2250082
 nsources_mn |   .1298578   .0187345     6.93   0.000     .0931388    .1665767
nsources_de~a |    .028249   .0027888    10.13   0.000      .022783    .0337149
 govern_share |   .0010022    .000121     8.29   0.000     .0007652    .0012393
       _cons |   14.80668   .0817325   181.16   0.000     14.64648    14.96687
-------------+----------------------------------------------------------------
     sigma_u |  .90698522
     sigma_e |   .2821005
         rho |  .91179291   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The coefficients for `nsources_mn` and `nsources_delta` are equal to those estimated in the Between Effects and Fixed Effects models respectively.

Furthermore we can test whether the between and within effects are equal:

```
test nsources_mn = nsources_delta
```

```
 ( 1)  nsources_mn - nsources_delta = 0

           chi2(  1) =   28.78
         Prob > chi2 =    0.0000
```

An equivalent approach is to use the `mundlak` command:

```
mundlak linc orgage localc west genchar nsources govern_share, hybrid
```

```
The variable orgage does not vary sufficiently within groups and will not be use
> d to create additional regressors.
0% of the total variance in orgage is within groups.

The variable localc does not vary sufficiently within groups and will not be use
> d to create additional regressors.
0% of the total variance in localc is within groups.

The variable west does not vary sufficiently within groups and will not be used
> to create additional regressors.
0% of the total variance in west is within groups.

The variable genchar does not vary sufficiently within groups and will not be us
> ed to create additional regressors.
0% of the total variance in genchar is within groups.

+-----------------------------------------------+
|             Variable |     RE     |   Hybrid   |
|----------------------+------------+-----------|
|               orgage |      0.005 |      0.005 |
|               localc |     -0.332 |     -0.329 |
|                 west |      0.080 |      0.097 |
|              genchar |     -0.273 |     -0.295 |
|             nsources |      0.030 |            |
|         govern_share |      0.001 |            |
|       diff__nsources |            |      0.028 |
|   diff__govern_share |            |      0.001 |
|       mean__nsources |            |      0.134 |
|   mean__govern_share |            |      0.000 |
|                _cons |     15.160 |     14.797 |
|----------------------+------------+-----------|
|                    N |      23826 |      23826 |
|                  N_g |   2166.000 |   2166.000 |
|                g_min |     11.000 |     11.000 |
|                g_avg |     11.000 |     11.000 |
|                g_max |     11.000 |     11.000 |
|                  rho |      0.912 |      0.912 |
|                 rmse |      0.282 |      0.282 |
|                 chi2 |    507.117 |    537.048 |
|                    p |      0.000 |      0.000 |
|                 df_m |      6.000 |      8.000 |
|                sigma |      0.950 |      0.950 |
|              sigma_u |      0.907 |      0.907 |
|              sigma_e |      0.282 |      0.282 |
|                 r2_w |      0.014 |      0.014 |
|                 r2_o |      0.083 |      0.095 |
|                 r2_b |      0.089 |      0.102 |
+-----------------------------------------------+
```

Mundlak approach

Random Effects model assumes that observed and unobserved effects are uncorrelated - an often unrealistic assumption (Gayle and Lambert, 2018).

We can relax this assumption using the *Mundlak approach*, which works by including unit-level means for the time-varying explanatory variables in the Random Effects model.

```
bys regno: egen orgage_mn = mean(orgage)
bys regno: egen govern_share_mn = mean(govern_share)
```

```
xtreg linc orgage localc west genchar nsources govern_share ///
    govern_share_mn nsources_mn orgage_mn, re
est store mund
```

```
Random-effects GLS regression                   Number of obs     =     23,826
Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:
     within  = 0.0140                                         min =         11
     between = 0.1042                                         avg =       11.0
     overall = 0.0976                                         max =         11

                                                Wald chi2(9)      =     557.99
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |   .0069072   .0005802    11.90   0.000       .00577    .0080444
      localc |  -.3282906   .0414364    -7.92   0.000    -.4095044   -.2470768
        west |   .1167392   .0805284     1.45   0.147    -.0410935    .2745719
     genchar |  -.3210918   .0449127    -7.15   0.000    -.4091191   -.2330644
    nsources |   .0289886   .0027931    10.38   0.000     .0235142     .034463
govern_share |   .0010325   .0001225     8.43   0.000     .0007923    .0012726
govern_shar~n |  -.0007575   .0007578    -1.00   0.317    -.0022428    .0007277
 nsources_mn |    .109471   .0195752     5.59   0.000     .0711044    .1478376
   orgage_mn |  -.0034024   .0007524    -4.52   0.000    -.0048772   -.0019277
       _cons |   14.85058   .0835091   177.83   0.000     14.68691    15.01426
-------------+----------------------------------------------------------------
     sigma_u |  .90700183
     sigma_e |   .2821005
         rho |  .91179586   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
quietly xtreg linc orgage localc west genchar nsources govern_share, fe
est store fixed
```

```
est table fixed mund
```

```
----------------------------------------
    Variable |   fixed         mund
-------------+--------------------------
      orgage |   .0069072     .0069072
      localc |  (omitted)    -.3282906
        west |  (omitted)     .11673923
     genchar |  (omitted)    -.32109178
    nsources |  .02898861     .02898861
govern_share |  .00103247     .00103247
govern_sha~n |               -.00075753
 nsources_mn |                .10947099
   orgage_mn |               -.00340245
       _cons |  14.715042     14.850581
----------------------------------------
```

```
quietly xtreg linc orgage localc west genchar nsources govern_share ///
    govern_share_mn nsources_mn orgage_mn, re

test govern_share_mn = nsources_mn = orgage_mn
```

```
 ( 1)  govern_share_mn - nsources_mn = 0
 ( 2)  govern_share_mn - orgage_mn = 0

           chi2(  2) =   43.79
         Prob > chi2 =    0.0000
```

The *Mundlak approach* is an alternative to the *Hausman test*.

Dynamic panel models

The models are suitable for when you have repeated contacts data and your (lagged) outcome variable serves also serves as one of your explanatory variables.

The inclusion of lagged outcome variables poses as an issue as the lagged variables are possibly correlated with the unobserved effects (Gayle and Lambert, 2018).

```
use "./data/charity-panel-analysis-2020-09-10.dta", clear
xtset regno fin_year
```

```
(Contains annual accounts of charities in E&W for financial years 2006-2017)

       panel variable:  regno (strongly balanced)
        time variable:  fin_year, 1 to 11
                delta:  1 unit
```

```
capture gen linc_lag = L.linc
l regno fin_year linc linc_lag in 1/22, clean
```

```
        regno   fin_year       linc   linc_lag  
  1.   200048    2006-07   14.00189          .  
  2.   200048    2007-08   14.17788   14.00189  
  3.   200048    2008-09    14.1851   14.17788  
  4.   200048    2009-10    14.2326    14.1851  
  5.   200048    2010-11    14.1709    14.2326  
  6.   200048    2011-12   14.14801    14.1709  
  7.   200048    2012-13     14.376   14.14801  
  8.   200048    2013-14   14.29996     14.376  
  9.   200048    2014-15   14.26031   14.29996  
 10.   200048    2015-16   14.30113   14.26031  
 11.   200048    2016-17   14.37021   14.30113  
 12.   200051    2006-07     17.664          .  
 13.   200051    2007-08   17.60568     17.664  
 14.   200051    2008-09   17.44065   17.60568  
 15.   200051    2009-10   16.46766   17.44065  
 16.   200051    2010-11   16.32526   16.46766  
 17.   200051    2011-12    16.4079   16.32526  
 18.   200051    2012-13   16.35779    16.4079  
 19.   200051    2013-14   16.04346   16.35779  
 20.   200051    2014-15   15.71779   16.04346  
 21.   200051    2015-16   15.42241   15.71779  
 22.   200051    2016-17   15.51123   15.42241  
```

```
xtreg linc orgage localc west genchar nsources govern_share linc_lag, re
```

```
Random-effects GLS regression                   Number of obs     =     21,660
Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:
     within  = 0.2673                                         min =         10
     between = 0.9963                                         avg =       10.0
     overall = 0.9320                                         max =         10

                                                Wald chi2(7)      =  296585.19
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |  -.0000262   .0000438    -0.60   0.549     -.000112    .0000596
      localc |  -.0065562   .0038027    -1.72   0.085    -.0140093    .0008968
        west |   .0009973    .007296     0.14   0.891    -.0133026    .0152973
     genchar |   -.005033   .0040519    -1.24   0.214    -.0129746    .0029087
    nsources |   .0102149   .0014779     6.91   0.000     .0073183    .0131116
govern_share |  -.0001568   .0000584    -2.69   0.007    -.0002712   -.0000424
    linc_lag |   .9719555    .001881   516.72   0.000     .9682687    .9756422
       _cons |   .4086979    .028964    14.11   0.000     .3519294    .4654663
-------------+----------------------------------------------------------------
     sigma_u |          0
     sigma_e |  .23554516
         rho |          0   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Note how large the coefficient is for the lagged variable (and how much smaller the others have become). This is a common issue when including lagged outcome variables as one of the explanatory variables i.e., the lagged variable soaks up all of the variation accounted for by the unobserved unit-specific effects.

```
xtreg linc orgage localc west genchar nsources govern_share linc_lag, fe
```

```
note: localc omitted because of collinearity
note: west omitted because of collinearity
note: genchar omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =     21,660
Group variable: regno                           Number of groups  =      2,166

R-sq:                                           Obs per group:
     within  = 0.2705                                         min =         10
     between = 0.9695                                         avg =       10.0
     overall = 0.9098                                         max =         10

                                                F(4,19490)        =    1807.17
corr(u_i, Xb)  = 0.9009                         Prob > F          =     0.0000

------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      orgage |   .0018829   .0005609     3.36   0.001     .0007836    .0029822
      localc |          0  (omitted)
        west |          0  (omitted)
     genchar |          0  (omitted)
    nsources |   .0211613   .0024664     8.58   0.000     .0163271    .0259956
govern_share |   .0005565   .0001088     5.12   0.000     .0003433    .0007698
    linc_lag |   .5167688   .0062088    83.23   0.000     .5045989    .5289386
       _cons |   7.147677   .0953486    74.96   0.000     6.960786    7.334568
-------------+----------------------------------------------------------------
     sigma_u |  .46351877
     sigma_e |  .23554516
         rho |  .79476462   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(2165, 19490) = 3.29                 Prob > F = 0.0000
```

A set of dynamic panel models — commonly known as *Arrelano-Bond* models — have been developed to address the inclusion of a lagged outcome as an explanatory variable.

They also have the advantage of controlling for "initial conditions".

That is, data collection sometimes interrupts an ongoing social process, and thus the outcome observed at the first time point is partially accounted for factors not measured at first time point (Gayle and Lambert, 2018).

Latent growth curve models

Statistical modelling of repeated contacts data.

Focuses on trajectory, trend or growth in an outcome over time **within** units.

And how these trajectories are linked to observed and unobserved differences **between** units.

Latent growth curve models can be estimated using a *Multilevel modelling* framework — random intercepts, random slopes.

They can also be estimated using a *Structural Equation Modelling (SEM)* framework — there exists underlying continuous trajectory of change that is not directly observed.

Honesty time

[Not an area I know a great deal about - see the reading list for suggested resources]
