Thinking about Cross-Sectional and Longitudinal Data

Dr Vernon Gayle

Longitudinal data analysis is important because it permits insights into the processes of change. Davies (1994) states this glib claim is inadequate and certainly fails to convince many social science researchers who are concerned with substantive rather than methodological challenges. What is required is an understanding of the limitations of cross-sectional analysis.

This handout outlines four central issues regarding longitudinal data analysis. These are **age and cohort effects**, **direction of causality**, **state dependence** and **residual heterogeneity**. These issues are complicated and only a cursory explanation is given in this handout!

Age & Cohort Effects

Table 1 shows some data on the average number of days spent off of the road for a particular make and model of car. It appears that older cars, on average, spend more days per year off the road. Buying one of these cars new looks like a bad idea because they appear to 'age' badly. There appears to be an 'ageing effect'.

Table 1 Cross-sectional Data on Car Reliability and Age

Age of car (years)	1	2	3	4	5
Average number of days off the road	3	4	15	16	18

However, the manufacturer tells me that cars made more recently are better. For example the ones made in the last two years are more reliable and only spend 3 or 4 days of off the road on average. The manufacturer suggests that there is a 'cohort effect'. The cohort of cars made 3 –5 years ago was less reliable. Could this be true?

Without longitudinal data we cannot unravel whether age or cohort (or a combination of each) provides the correct explanation. Below in Table 2 and Table 3 are some longitudinal data on reliability for two cars, the Ford Viva and the Vauxhall Capri. If you were buying a new car, which would you choose on reliability?

Table 2 Longitudinal Data on Car Reliability and Age (Ford Viva)

	Year of Manufacture				
	1997	1998	1999	2000	2001
Age of Car (years)					
1	3	4	3	3	4
2	4	4	3	3	
3	10	10	10		
4	16	16			
5	18				

Table 3 Longitudinal Data on Car Reliability and Age (Vauxhall Capri)

	Year of Manufacture				
	1997	1998	1999	2000	2001
Age of Car (years)					
1	18	18	15	3	4
2	18	16	15	3	
3	18	16	15		
4	18	16			
5	18				

In Table 2 we can see a clear ‘ageing effect’. As the Ford Viva cars get older they become less reliable.
In Table 3 we can see a clear ‘cohort effect’. So far the Vauxhalls made in 2000 and 2001 appear to be much more reliable than the ones manufactured between 1997 and 1999.

The Vauxhall Capri appears to be a better purchase in terms of reliability than the Ford Viva.

Cross-sectional data are uninformative about age and/or cohort effects – to untangle these effects we need longitudinal data.

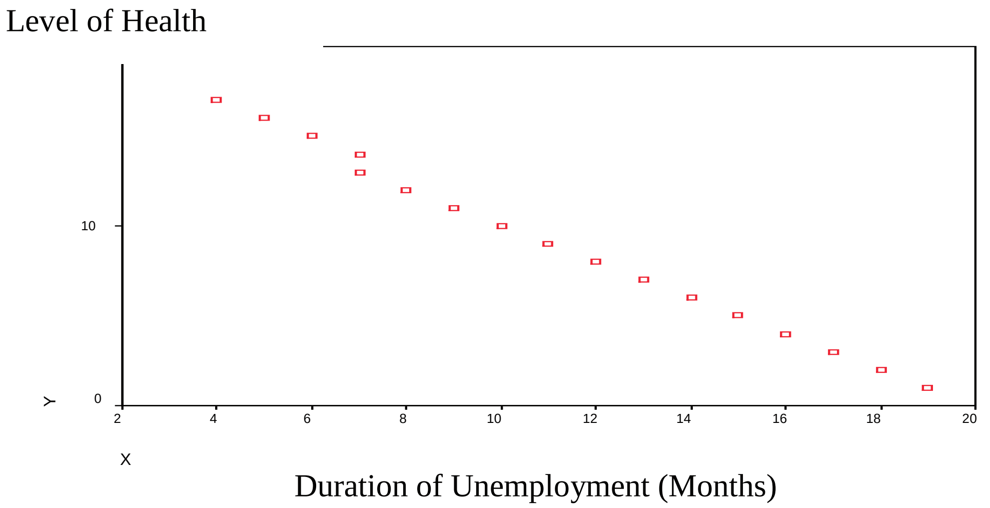
Direction of Causality

There is unequivocal evidence from cross-sectional data that, overall, the unemployed have poorer health. This is consistent with two hypotheses; 1) unemployment causes ill health 2) ill health causes unemployment.

Imagine that I had a survey that collected data on people’s employment status (e.g. whether or not they were unemployed) and also on their level of health. It would also be plausible that we might ask people how long they had been unemployed. Generally, we might expect to find a negative relationship. Those that had been unemployed for longer would have lower levels of health.

The cross-sectional data reported in Figure 1 support the hypothesis that unemployment causes ill health (ill health is measured on a 20 point scale; lower score indicating poorer health). The people that have been unemployed for longer periods have lower levels of health.

Figure 1 Plot of Level of Health and Duration of Unemployment – Cross-sectional Data



However, these cross-sectional data are also consistent with the second hypothesis; that ill health causes unemployment. If ill health prevents a person from working those with less severe ill health will recover and return to work. With the increasing duration of unemployment those with less severe ill health will be progressively under-represented, while those with more severe ill health will the over represented. This is known as 'sample selection bias'. This sample selection bias could therefore explain the cross-sectional picture of declining health with duration of unemployment.

Without longitudinal data we cannot solve this puzzle. To get to grips with this we would need longitudinal data. Tables 4 and 5 report (hypothetical) data on the level of health (measured on a 20 point scale) and employment status for two individuals. **This is a simplified depiction of the issue.**

Table 4 Level of Health and Employment Status (Individual A)

Month	Level of Health	Employment Status
1	17	Employed
2	17	Employed
3	17	Employed
4	17	Unemployed
5	17	Unemployed
6	10	Unemployed
7	6	Unemployed
8	5	Unemployed
9	4	Unemployed
10	3	Unemployed
11	2	Unemployed
12	1	Unemployed

These data in Table 4 (Individual A) suggest that unemployment has preceded ill health and as the duration of their unemployment increased their level of health declined. This is consistent with the idea that unemployment leads to ill health.

Table 5 Level of Health and Employment Status (Individual B)

Month	Level of Health	Employment Status
1	17	Employed
2	1	Employed
3	1	Employed
4	1	Unemployed
5	1	Unemployed
6	1	Unemployed
7	1	Unemployed
8	1	Unemployed
9	1	Unemployed
10	1	Unemployed
11	1	Unemployed
12	1	Unemployed

These data in Table 5 (Individual B) suggest that ill health preceded unemployment. This is consistent with the idea that ill health causes unemployment.
Note: if we had undertaken a cross-section survey in month 12 for these two individuals

* A would have been unemployed for 9 months and have a health score of 1
* B would have been unemployed for 9 months and have a health score of 1

However, we can see that with longitudinal data we can start to untangle this.

*If more individuals experienced the same relationship between health and employment as Individual A, this would lend support to the hypothesis that unemployment causes to ill health.

*If more individuals experienced the same relationship between health and employment as Individual B, this would lend support to the hypothesis that ill health causes unemployment.

State Dependence

This is the idea that current behaviour is influenced by past or previous behaviour. Consider the following two examples.

*The chances that you are married in any given month are highly contingent on whether or not you are married in the previous month. Most people don't change their marital status with great frequency.

*Your chances of being employed in any month are contingent on whether or not you were employed in the previous month. The labour market tends not to be sufficiently volatile to cause the switching of employment states.

Therefore we can see why we might want to take account of prior information when examining current situations. This is only possible with longitudinal information.

Residual Heterogeneity

This is a complicated issue. Put simply, longitudinal data allows us to increase control in our analysis for residual heterogeneity. Residual heterogeneity is a term that refers to the omission of explanatory variables in our analysis. These omitted explanatory variables are either unmeasured or un-measurable.

If you want to understand this more fully…
Cross-sectional data allows us to undertake analysis between cases (usually individuals). The advantage of longitudinal data is that, in addition to between cases analysis, it allows us to undertake analysis within cases because we have measures for the same individual at different time points.
In our dataset there might be individuals with similar characteristics but they behave differently at different time points. This would suggest that some of their behaviour might be explained by unmeasured (and possibly un-measurable) variables (e.g. motivation). The possibility of substantial variation due to unmeasured and possibly un-measurable variables is known as 'residual heterogeneity'.

References and Further Reading

Davies, R.B. (1994) 'From Cross-Sectional to Longitudinal Analysis', in Dale, A. and Davies, R.B. *Analyzing Social & Political Change: A Casebook of Methods*, Sage.

Plewis, I. (1985) *Analysing Change: Measurement and Explanation Using Longitudinal Data*, Wiley.

Ruspini, E. (2000) 'Longitudinal Research in the Social Sciences', *Social Research Update,* 28. <http://www.soc.surrey.ac.uk/search/search.htm>

Taris, T. (2000) *Longitudinal Data Analysis*, Sage.