ROBIN HENDERSON

Missing Data in Longitudinal Studies

Introduction

In his January 2006 presentation to the Centre for Advanced Study, Professor Jeremy Hawthorn discussed the variety of narrative styles provided by Holocaust survivors recounting their experiences. Part of the subsequent discussion focussed on the observation that only the survivors can tell of what happened, and it could be that in some way the survivors are different from the large number of people who did not survive. In his book *This Way for the Gas, Ladies and Gentlemen*, survivor Tadeusz Borowski makes this point explicitly, in that the question "But how did it happen that you survived?" is important to those with his experience, although it is impossible to answer. Was he more healthy and stronger than others? Or was he more selfish and cunning? Or was it just random chance?

Professor Robin Henderson
School of Mathematics and Statistics,
University of Newcastle upon Tyne, UK
E-mail:
Robin.Henderson@newcastle.ac.uk
CAS Fellow 2005/2006

In statistical terms, this is an example of a *selection effect* for longitudinal or event history data. Statistical methods are used to draw inferences from samples and of course this means that the sample needs to be representative of the population in question. This is straightforward for cross-sectional data, but the picture becomes more blurred whenever there is longitudinal follow-up, i.e. where people are followed over time to see how responses change. There is then the possibility of losing contact with people, which can happen for a variety of reasons. If the study is of the elderly or very ill, then a number may die before the follow-up period is complete. If the study is a clinical trial of a new treatment, then some patients may withdraw as a result of side-effects. Or people may choose not to visit clinics since they consider themselves healthy. Others may be too ill to attend out-patient clinics, and so on. In all of these circumstances, the people who complete the study will be a subset of those who began it, and no matter how careful we are to ensure that our original sample is representative, there is no guarantee that the same will be true of our closing sample.

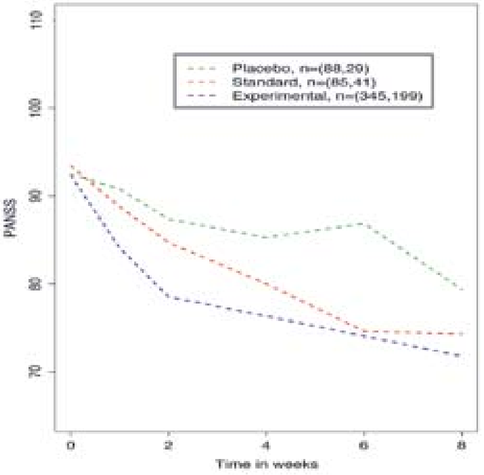

Figure 1. Observed means for schizophrenia data. The numbers in the legend give the group sizes at the beginning and end of the trial.

Figure 1 provides a concrete example, using data from a clinical trial on the treatment of schizophrenia. The graph is based on results from 518 patients who began the trial, of whom 88 received placebo, 85 were given a standard treatment, and 345 were given a new experimental treatment, at one of four dosage levels. For the sake of simplicity, the dose is not considered here. A measure of mental health, the positive and negative symptom scale (PANSS), was then obtained at weeks 0, 1, 2, 4, 6, and 8 of the trial. On this scale, high values reflect worse condition. Unfortunately, a large number of patients dropped out before completion for a variety of reasons. Within the experimental treatment group 58 per cent of patients participated in the whole trial, 48 per cent in the standard treatment group and 33 per cent in the placebo group.

The plot shows the mean PANSS score at each point in time for the patients remaining in the trial. Within each group, there is evidence of a decline in the mean score over time and hence an improvement in mental health. This could be real, reflecting benefits of treatment. Improvement even in the placebo group might be plausible since recruitment into the trial may upset fragile people and their original scores may have been uncharacteristically high. On the other hand, it could be a selection effect as described above, as we could argue that the most ill patients tend to drop out, meaning that the average mental health of the continuers is better. Thus the observed drop in the mean may be an artefact since we do not compare identical groups at all time points. The people who make it to the end of the study may not be representative of those who began it.

Rubin’s taxonomy for missing data

We will return to the schizophrenia data later, but first it may be useful to describe how missingness mechanisms can be classified. For this, we will follow the standard taxonomy (e.g. Rubin 1976) using a numerical example for illustration.

Suppose that measurements are scheduled at week 1 and week 2 of a trial. We will call these Y_1 and Y_2 and we assume that each can be either 0 or 1. We begin with 200 patients, half of whom have $Y_1 = 0$ and half $Y_1 = 1$. These are all observed at week 1 but unfortunately a number of patients drop out before their week 2 measurement is obtained. How can we estimate the mean response at week 2? Table 1 gives three scenarios to consider. In each case, we give a 2 × 2 table of outcomes for the people who are observed on both occasions (the ‘completers’) and for the true but unobserved outcomes for the others (the ‘dropouts’). The final part of each block summarises the data actually observed.

In Block A, the probability of dropout is 0.5 for all individuals, irrespective of their Y_1 and Y_2 values. This is the so-called *missing completely at random* (MCAR) situation, which causes no problems for the analysis. On inspection of the table, we see that a total of 140 people from the 200 have $Y_2 = 1$, observed or not, and so the true mean is 0.7. If we consider only the 100 people observed at week 2 we find 70 have $Y_2 = 1$, and so the observed mean is 0.7, the correct value.

It would be nice if things were always that simple but the problem becomes murkier when we move to Block B. Here, people with $Y_1 = 0$ have probability 0.8 of dropping out but those with $Y_1 = 1$ have probability 0.2 of dropping out. In both cases, the value of Y_2 does not affect dropout. This is an example of data *missing at random* (MAR), where the dropout probability can depend on quantities which are always observed

A.

Completers

		Y_2		
		0	1	
Y_1	0	25	25	50
	1	5	45	50
		30	70	

Dropouts

		Y_2		
		0	1	
Y_1	0	25	25	50
	1	5	45	50
		30	70	

Observed Data

$Y1 = 0$	no Y_2	50
$Y1 = 1$	no Y_2	50
$Y1 = 0$	$Y_2 = 0$	25
$Y1 = 0$	$Y_2 = 1$	25
$Y1 = 1$	$Y_2 = 0$	5
$Y1 = 1$	$Y_2 = 1$	45

B.

Completers

		Y_2		
		0	1	
Y_1	0	10	10	20
	1	8	72	80
		18	82	

Dropouts

		Y_2		
		0	1	
Y_1	0	40	40	80
	1	2	18	20
		42	58	

Observed Data

$Y1 = 0$	no Y_2	80
$Y1 = 1$	no Y_2	20
$Y1 = 0$	$Y_2 = 0$	10
$Y1 = 0$	$Y_2 = 1$	10
$Y1 = 1$	$Y_2 = 0$	8
$Y1 = 1$	$Y_2 = 1$	72

C.

Completers

		Y_2		
		0	1	
Y_1	0	10	40	50
	1	2	72	74
		12	112	

Dropouts

		Y_2		
		0	1	
Y_1	0	40	10	50
	1	8	18	26
		48	28	

Observed Data

$Y1 = 0$	no Y_2	50
$Y1 = 1$	no Y_2	26
$Y1 = 0$	$Y_2 = 0$	10
$Y1 = 0$	$Y_2 = 1$	40
$Y1 = 1$	$Y_2 = 0$	2
$Y1 = 1$	$Y_2 = 1$	72

Table 1. Some dropout patterns.

(Y_1) but not on observations which may be missing (Y_2). Examining the table, we see 100 people have observed Y_2 but their mean is 0.82 instead of the true 0.7. Very simple estimation methods no longer work. Fortunately, there are a variety of solutions, one of which uses the so-called inverse probability of observation weighting.

This is a method familiar in social surveys for over 50 years but popularised for longitudinal analysis by Professor Jamie Robins and colleagues (e.g. Robins et al. 1995). The idea is to take the fully observed patients but give more weight in the analysis to those who we think are more rarely observed, to let them stand as a type of proxy for unobserved people. Thus in Block B we have an 80 per cent dropout rate for those with $Y_1 = 0$ or, with a change of scale, an observation rate of only one in five. This means we assign the observed people at week 2 a weight of five if they have $Y_1 = 0$: one for themselves and four for missing people. For $Y_1 = 1$, the observation rate is 4/5 and those people are assigned a weight of 5/4. Our estimated week 2 mean is then

$$(5 \times 10 + 5 \times 72/4)/200 = 0.7$$

as required.

The final category is *missing not at random* (MNAR) or, more elegantly, *informative dropout*. This is illustrated in Block C, where the dropout rate is 80 per cent for those with $Y_2 = 0$ and 20 per cent for those with $Y_2 = 1$, irrespective of Y_1. Both the simple and inverse probability methods would give the wrong answer in this case, and we can make progress only by making further assumptions about how the data are generated.

Joint analysis of longitudinal and event history data using random effects

Moving beyond the simple example, suppose now that we are interested in the development over time of a measure of the health of an individual, such as kidney performance after a transplant operation. This is not observable and instead snapshots, Y_1, Y_2,..., are obtained at times as determined by the experimenter. The snapshots may be subject to measurement error or may only indirectly measure the true state, such as creatinine levels as markers for kidney function. Together, the Y measurements provide a longitudinal profile for each subject.

Alongside the measurement schedule, we assume there is a sequence of event history data S: stochastic points in time whose intensity may also be related to the underlying value of the subject's 'true health'. These could be recurrent events such as rejection episodes for a kidney transplant recipient, or a single event whose occurrence terminates observation of the measurement sequence, such as kidney failure or the death of the patient.

The purpose of the analysis is to determine the relationship between Y and S and how both are influenced by factors such as age, sex, treatment and so on. Adjusting for selection effects as described in the previous sections is a special case of this problem, where our interest is primarily in the Y measurements and dropout times (S) are a nuisance. Another special case has the alternate priorities, with the main interest focussed on the event processes S, with Y being a time-varying and error-prone marker of underlying intensity. This is extremely important given the explosion of recent interest in biomarkers, which are measures of disease progression, sometimes obtained at the cellular level. How well do these predict subsequent outcomes, such as longevity?

Hogan et al. (2004) provide a nice overview of the methods available, most of which are rather complicated. We will mention briefly just one class of methods, a *random effects* approach, where the interesting association between the Y and S components is obtained by their shared dependence on an unobserved patient-specific quantity W, a so-called random effect. We then make modelling assumptions about W and use these with the data to estimate parameters of interest.

A simple illustration will close this section. Referring to the schizophrenia data, we assume that individuals have their own intrinsic mental health W, which we assume has a normal distribution. People with high W have high PANSS scores, and people with low W have low PANSS scores. We then assume that the chance of dropping out also depends on W. Since W is not observed, we have to average over its possible values, using modern computationally intensive methods. This is feasible and leads to results like those given in Figure 2, which adjusts the PANSS means to account for selection effects. In this example we see that the mean stays roughly constant in the placebo group, that there is improvement in the standard treatment group but more improvement under the new treatment. These suggestions were in fact confirmed by formal statistical testing (Henderson et al. 2000).

A cautionary comment

Methods for dealing with missing data are now widely available and the field is maturing quickly. However, almost all methods rely on making assumptions about the mechanism that causes the data to have missing values. Unfortunately, these assumptions are invariably not testable based on observed data. We may see strong positive correlation between successive observed values, for instance, and it is tempting to assume that missing ones would follow a similar pattern. (In fact this is the basis for most of the methods.) But it is quite possible that an observation is missing precisely because it *did not* follow the regular pattern: something went wrong and that caused the dropout. We can never tell for sure. As a failsafe, *sensitivity analyses* are recommended to assess the robustness of conclusions across a variety of models, methods and assumptions. These are not perfect and it is always useful to bear in mind that the best way to deal with missing data is not to have them in the first place.

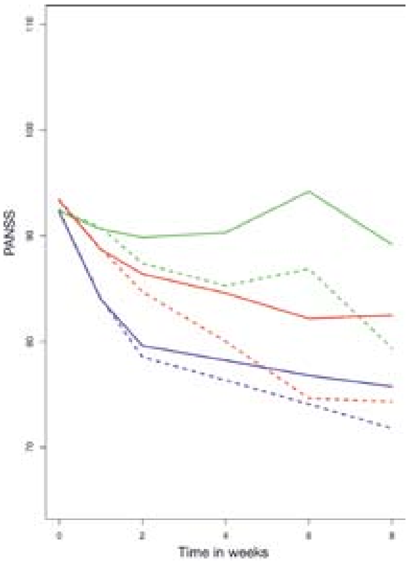

Figure 2. Observed (dotted) and adjusted (solid) means for schizophrenia data.

Acknowledgement

I am extremely grateful for the opportunity to spend time at the Centre for Advanced Study in Oslo, and for the kindness and consideration of all those I met there.

References

Aalen, O.O. “Effects of frailty in survival analysis”. *Statistical Methods in Medical Research*, 3, 1994, 227-243.

Borowski, T. *This Way for the Gas, Ladies and Gentlemen*. Penguin, 1992, London.

Henderson, R., Diggle, P.J. and Dobson, A. “Joint modelling of longitudinal measurements and event time data”. *Biostatistics*, 2000,1, 465-480.

Hogan, J.W., Roy, J. and Korkontzelou, C. “Tutorial in biostatistics – handling drop-out in longitudinal studies”. *Statistics in Medicine*, 23, 2004, 1455–1497.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. “Analysis of semiparametric regression models for repeated outcomes in the presence of missing data”. *Journal of the American Statistical Association*, 90, 1995 106–121.

Rubin, D.B. “Inference and missing data”. *Biometrika,* 63, 1976, 581–592.