CENTRE FOR APPLIED ECONOMIC RESEARCH

WORKING PAPER
(2005/06)

Optimal recall length in survey design

By Philip M. Clarke, Denzil G. Fiebig and Ulf-G Gerdtham

ISSN 13 29 12
ISBN 0 7334 2296 9
www.caer.unsw.edu.au

Faculty of Commerce and Economics

UNSW
THE UNIVERSITY OF NEW SOUTH WALES

Optimal recall length in survey design*

Philip M. Clarke[1,2], Denzil G. Fiebig[3] and Ulf-G Gerdtham[4]

[1]Health Economic Research Centre, University of Oxford, UK
[2]Dept of Endocrinology Prince of Wales Hospital, Australia
[3]School of Economics, University of New South Wales, Australia
[4]Department of Clinical Sciences, Lund University, Sweden

Abstract:

Self-reported data collected via surveys are a key input into a wide range of research conducted by economists. It is well know that such data are subject to measurement error that arises when respondents are asked to recall past utilisation. Survey designers must determine the length of the recall period and face a trade-off as increasing the recall period provides more information, but increases the likelihood of recall error. A statistical framework is used to explore this trade-off. Finally we illustrate how optimal recall periods can be estimated using hospital use data from Sweden's Survey of Living Conditions.

Classification codes: C42; I1

6 October 2005

* This research was partially supported by the NHMRC through a Program Grant No. 254202 (Fiebig) and a Project Grant No. 300565 (Clarke). Financial support from the Swedish Council for Working Life and Social Research (2002-0376) and the Swedish National Institute of Public Health is gratefully acknowledged (Gerdtham). Comments by participants at the 2005 Australian Health Economics Society Conference held in Auckland, especially those by Jim Butler, are gratefully acknowledged.

1. Introduction

Self-reported data collected via surveys are a key input into a wide range of research conducted by economists and other social scientists. For example, in health economics economic evaluations of health care interventions are often based on self-reported health care data that is collected during the course of a study. It is also routinely collected in national health surveys such as the Health Survey for England, Australia's National Health Survey or the Medical Expenditure Survey in the US. While these surveys involve interviews conducted over the course of a year, there is considerable variation in the period over which subjects are asked to recall their previous health care use. This has been highlighted by a recent OCED study comparing inequity in access to health care across 21 developed countries based on health surveys in different countries. The period of recall for primary care and aspects of hospital use ranged from 2 weeks in the Australian National Health Survey to one year in the European Community Household Panel (Health Equity Research Group, 2004).

It has been widely recognized that there is an inverse relationship between the length of time over which subjects are asked to recall prior use and the accuracy of the reported estimates. The longer the period of recall the greater the likelihood of error. An early study by Sudman and Bradburn (1973) on the impact of the length of the recall period on response suggested two types of memory error may arise: (i) a respondent may forget an episode of use entirely; and (ii) they may remember the episode but incorrectly recall when it occurred. In regard to the latter, a particular form of error is *telescoping* which arises when the respondent recalls events that occurred before the period in question. For example, a person my report they have been in hospital in the last three months, even though the actual event was five months prior to the date of interview. While these errors can work in opposite directions, recent reviews of validation studies by Bound, Brown and Mathiowetz (2000) and Evans and Crawford (1999) that compared reported with actual use of different types of health care, suggest that under-reporting is more likely than over-reporting, especially in primary care. This potentially has implications for economic evaluation, particularly if there are differing degrees of recall between interventions.

While recall error is undesirable, a survey with a short length of recall provides very little information about an individual's normal health care use. As Deaton (1997, p. 24) has noted in the context of measuring general household consumption expenditure:

> "…if the object of the exercise is to estimate average consumption over a year, one extreme is to approach a sample of households on January 1 and ask each to recall expenditures for the last year. The other extreme is to divide the sample over the days of the year, and to ask each to report consumption for the previous day. The first method would yield a good picture of each household's consumption, but runs the risk of measurement error… [t]he second method… will give a good estimate of the mean consumption over all households… [but] will yield estimates of individual expenditure that… are only weakly related to normal expenditures…"

Suppose self-reported consumption or utilisation data are to be collected by a survey. At the survey design stage there is a choice of the length of the recall window. Even if the variable of interest is consumption over a particular period, say a year, errors induced by having long recall windows equal to the period of interest suggest that one could opt for a shorter window or sub-period as the basis of the survey question. While this provides a less error ridden measure it comes at the cost of providing less information. Hence, the basic research question is: "What is the optimal recall window over which to ask the utilisation question when recall error increases with window length but a longer window is likely to provide more (albeit imperfect) information directly relevant to the variable of interest?"

We investigate these tradeoffs in order to determine what factors impact on the optimal recall window over which to ask consumption and utilisation questions when recall error increases with window length, but a longer window has the advantage of providing more (albeit imperfect) information directly relevant to the variable of interest. A simple framework is used to capture the key dimensions of the research question and this analysis is supplemented by an illustration drawn from hospital usage data from Statistics Sweden's Survey of Living Conditions.

2. A framework for analysis

Denote the variable of interest for the i^{th} individual by Y_i. If this could be recalled without error then there would be no problem. Such a situation represents a

benchmark for comparisons as specific forms of recall errors are introduced. Suppose the period over which the utilisation variable is required, call this the target period, can be divided into S sub-periods where the division is determined so that individuals can accurately recall their utilisation over the most recent sub-period. If we were interested in how much someone spent on medications over a year but the longest window over which they can recall last period's drug expenditure without error is say a month then S is set to 12. Longer recall windows, denoted by w =1, 2, 3, … , S , are considered because monthly consumption may not be very representative of annual consumption. Moreover, we are ultimately interested in annual consumption and hence windows of less than a year imply that an imputation process must be undertaken in order to provide a predicted annual amount.

Now denote actual utilisation over the recall window w by Y_i^w. However, when asking the survey question the error ridden response will be given by:

$$(1) \quad X_i^w = Y_i^w + v_i^w$$

where v_i^w represents the measurement error. By assumption, $X_i^1 \equiv Y_i^1$ but for $w > 1$ measurement error can occur because of incomplete or inaccurate recall. There are other potential causes for measurement error. In particular, respondents could act strategically or simply provide false answers to sensitive questions. In such cases the form of the measurement error is likely to be different from that associated with recall problems as are the likely remedies; see for example Carson, Groves, and Machina (1999) and Philipson and Malani (1999). In what follows, we concentrate on measurement errors that occur because of difficulties in recalling past events.

Formalization of the design problem will depend on what the ultimate objectives are of the analyses to be undertaken using these data. Suppose interest centres on characterizing the variable of interest and hence the statistical problem is to estimate $E(Y_i^S)$, the mean utilisation over the target period. Given a sample of N individuals and a window of w, an obvious estimator would be a suitably weighted sample mean with the weights defined to scale up the sub-period utilisation to a target period estimate. Define this estimator as:

$$(2) \quad \overline{X}_w = N^{-1} \sum_{i=1}^{N} \left(\frac{S}{w} \right) X_i^w .$$

For the present, assume that the relationship between sub-period utilisation and that in the target period is known. For example, to estimate the annual drug use of people with a chronic disease such as diabetes it may be sufficient to ascertain monthly use if their prescriptions do not vary over the course of a year. In practise, for many types of health care this is not likely to be the case and hence would represent an additional source of information loss associated with choosing $w<S$.

One possible objective could be to consider the impact of w on the bias in the estimation of the mean utilisation over the target period. For example, Benitez-Silva et al (2004) introduce the hypothesis of *rational unbiased reporting* as part of a framework for testing the validity of self-reported health measures. For our purposes this is unsatisfactory because unbiasedness alone does not capture the tradeoffs that are involved. Instead, we consider quadratic loss so that the survey design problem requires choice of w to minimize the $MSE(\overline{X}_w)$.

3. Some results for optimal recall windows

Further investigation requires more structure for the data generating process assumed to be underlying this problem. In what follows both discrete and continuous utilisation variables are considered. The issues that arise and some of the conclusions are different enough to justify this distinction.

3.1 Continuous case

The starting point for our analysis is the case where Y_i is assumed to be a continuous random variable. Assume that utilisation in each of the sub-periods are iid with mean and variance given by:

$$(3) \quad E(Y_i^w) = \frac{w}{S}\mu; \quad Var(Y_i^w) = \frac{w}{S}\sigma^2;$$

which implies $Y_i \equiv Y_i^S$ has mean and variance given by:

(4) $E(Y_i) = \mu; \quad Var(Y_i) = \sigma^2.$

For the measurement model, make the classical errors-in-variables assumption that v_i^w is independent of Y_i^w but allow

(5) $E(v_i^w) = h^*(w) = h(w)\left(\frac{w}{S}\right)\mu$

and

(6) $\text{var}(v_i^w) = g^*(w) = g(w)\left(\frac{w}{S}\right)\sigma^2.$

Thus, the mean and variance of the measurement errors are assumed to depend on the recall window and the $h(.)$ and $g(.)$ functions have been scaled to make these moments proportional to the mean and variance of the actual utilisation, Y_i^w. These functions introduce two dimensions to recall error. Non-zero values for $h(.)$ represent recall bias while a non-zero $g(.)$ implies more noisy measurements. Because no recall error is assumed when $w = 1$ it must be that $h(1) = g(1) = 0$. Further, assume both functions are monotonic over the interval 1 to S. $h(.)$ could either be monotonically deceasing if there was under reporting or monotonically increasing in the case of over reporting. $g(.)$ is expected to be monotonically increasing.

With this framework, the following can be derived:

$$E(\overline{X}_w) = N^{-1}\left(\frac{S}{w}\right)\sum_{i=1}^{N} E(X_i^w) = N^{-1}\left(\frac{S}{w}\right)\sum_{i=1}^{N} \mu\left(\frac{w}{S}\right)\left[1 + h(w)\right] = \mu\left[1 + h(w)\right]$$

and

$$\text{Var}(\overline{X}_w) = N^{-2}\left(\frac{S}{w}\right)^2 \sum_{i=1}^{N} Var(X_i^w) = N^{-2}\left(\frac{S}{w}\right)^2 N\left[1 + g(w)\right]\left(\frac{w}{S}\right)\sigma^2 = \left[1 + g(w)\right]\left(\frac{S}{w}\right)\left(\frac{\sigma^2}{N}\right)$$

leading to:

$$(7)\quad MSE(\overline{X}_w) = Var(\overline{X}_w) + \left[Bias(\overline{X}_w)\right]^2$$
$$= \left[1 + g(w)\right]\left(\frac{S}{w}\right)\left(\frac{\sigma^2}{N}\right) + \mu^2 h^2(w), \quad RMSE(\overline{X}_w) = \sqrt{MSE(\overline{X}_w)}$$

These expressions provide the basis for a number of comparisons that inform us on the question of how best to choose *w*. These comparisons are organized around several different cases.

Case 1: No recall bias

Here $h(w) = g(w) = 0$ and (7) simplifies to:

$$(8)\quad MSE(\overline{X}_w) = Var(\overline{Y}_w) = \left(\frac{S}{w}\right)\left(\frac{\sigma^2}{N}\right).$$

It is clear that the optimal recall window is $w = S$ as other choices produce less precise estimates. There is no gain from choosing a smaller recall window; such choices would involve collecting less data for the same sample size.

Case 2: Comparison of w =1 and w = S

A natural comparison is between using a recall window that involves no recall error ($w = 1$) and one where the question relates directly to the target period ($w = S$). In terms of our model structure, this case also avoids making additional assumptions about $h(.)$ and $g(.)$.

Given different parameter configurations, it is a simple matter to produce MSE comparisons but the main points will be made with one representative set of results. Table 1 provides relative root mean square error (RMSE) measures for a particular set of parameters chosen to be: N =1000, S = 12, $\mu = \sigma = 1$ and maximum values of $h(.)$ and $g(.)$ that are allowed to vary. Recall that these functions were scaled so as to represent proportions of the mean and variance of the actual utilisation. Consequently, unitary values for *hmax* and *gmax* imply substantial recall error in terms of either induced bias or added variability. Values in the table that are less than unity imply

that the recall window of $w = S$ is preferred over $w = 1$ if the objective is estimation of the population mean of utilisation over the target period.

When there is no recall bias, the results in the first row ($h(w) = 0$) of Table 1 indicate $w = S$ is strongly preferred over $w = 1$ even when the recall error induces substantial amounts of measurement noise, i.e. $g(w)$ is large. It is simple to demonstrate that with no recall bias the superiority of a window of S is even more pronounced if N and/or S are larger. Clearly it will not always be optimal to totally avoid recall error. This conclusion may still be valid even when recall bias is introduced. Consider the case where $hmax = 0.1$. Again it is better to ask about utilisation over the target period even though this involves both types of recall errors. The alternative of simply asking about utilisation for the most recent sub-period, while being measured without recall error, induces substantial variability in the resultant estimator because of substantial losses in information.

Table 1: Relative RMSE of alternative estimators of μ: Case 2 with N = 1000, S = 12, $\mu = \sigma = 1$*

	Gmax					
hmax	**0.0**	**0.2**	**0.4**	**0.6**	**0.8**	**1**
0.0	0.29	0.32	0.34	0.37	0.39	0.41
0.1	0.96	0.97	0.97	0.98	0.99	1.00
0.2	1.85	1.85	1.86	1.86	1.87	1.87
0.3	2.75	2.76	2.76	2.76	2.77	2.77
0.4	3.66	3.67	3.67	3.67	3.67	3.67
0.5	4.57	4.58	4.58	4.58	4.58	4.58

* Table entries represent $RMSE(\overline{X}_S) / RMSE(\overline{X}_1)$.

As more recall bias is introduced the preferred recall window switches to $w = 1$. Now the efficiency loss associated with this estimator is more than compensated by the large biases it avoids. A similar pattern can be shown to occur if μ is larger because again recall bias becomes more of a problem.

Case 3: Optimal w

Case 2 provided insights into the optimal recall window problem by comparing the extreme choices. Naturally it is possible that neither extreme is optimal and that the optimal recall window falls somewhere in between. Not surprisingly, whether this is

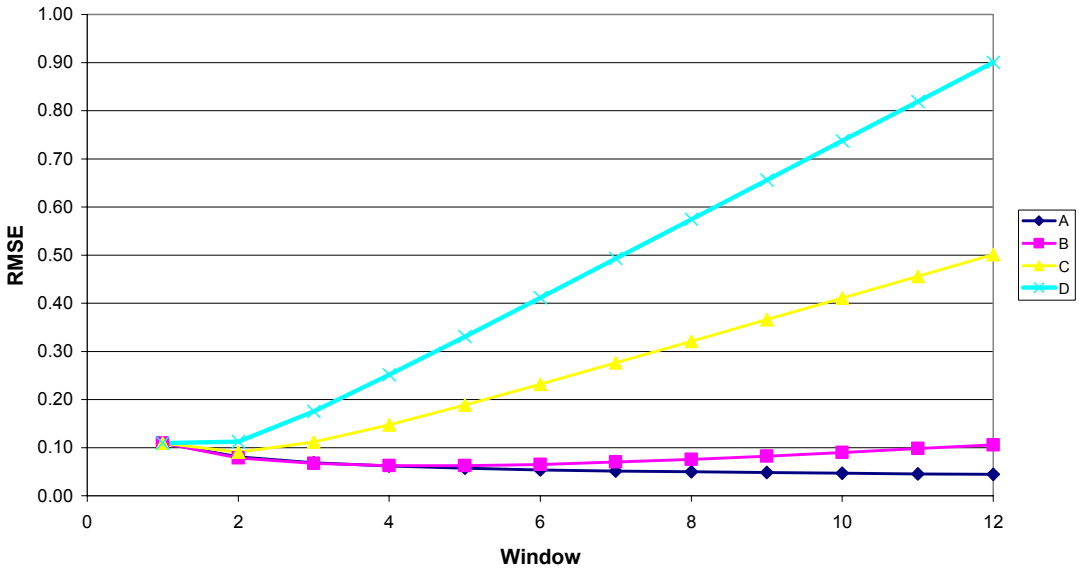
the case or not depends on parameter configurations and on assumptions made about the specific form of the $h(.)$ and $g(.)$ functions. In order to make some progress, first assume that these functions are both linear.

Figure 1 depicts a range of $RMSE(\overline{X}_w)$ functions that show the types of patterns that can result from different parameter configurations. All of the reported functions assume $N = 1000$, $S = 12$ and $\mu = \sigma = 1$ with the differences being generated by variable forms of recall error represented by setting alternative values for *hmax* and *gmax*. Start with the function labelled A where *hmax* = 0 and *gmax* = 1. This was covered under Case 2 where we concluded that $w = S$ was preferred to $w = 1$. Now we see that S is in fact the optimal choice of w as the RMSE function declines monotonically as w increases.

Functions B and C illustrate non-monotonic patterns. Function B was generated with *hmax* = 0.1 and *gmax* = 0.2. Again, from the discussion under Case 2, $w = S$ was found to be preferred to $w = 1$. However, it is possible to make a better choice for the recall window and the optimal w is in fact equal to 5. Similarly, function C with *hmax* = 0.5 and *gmax* = 0.2 yields a preference for $w = 1$ compared to $w = S$ but the optimal w is in fact equal to 2.

None of the parameter configurations considered under Case 2, together with the additional assumption of linear $g(.)$ and $h(.)$ functions, will generate a RMSE function that increases monotonically as w increases. However, such a result is possible if the recall bias is sufficiently large. Function D with *hmax* = 1.0 and *gmax* = 0 provides an example.

Figure 1: RMSE comparison



3.2 Discrete case

Many variables of interest will refer to incidence or participation and hence will be binary in nature. Have you visited your GP in the last 12 months? Have you had a Pap test in the last 2 years? If Y_i^w and X_i^w are binary variables then the population parameter of interest will be denoted by:

(9) $\Pr(Y_i^S = 1) = p$

and

(10) $\Pr(Y_i^w = 1) = \pi_w$.

Here measurement errors are defined as:

(11) $\pi_{10}(w) = \Pr(X_i^w = 1 \mid Y_i^w = 0)$, and
$\pi_{01}(w) = \Pr(X_i^w = 0 \mid Y_i^w = 1)$.

The form of the measurement errors represented by (11) makes it clear that the measurement error v_i^w defined by (1) will be negatively correlated with Y_i^w. Unlike the

continuous case, the classical errors-in-variable framework is not appropriate in the discrete case. [1]

Assume that the measurement errors depend on the recall window and that:

$$\pi_{10}(1) = \pi_{01}(1) = 0; \tag{12}$$

$$\frac{\partial \pi_{10}(w)}{\partial w} > 0, \frac{\partial \pi_{01}(w)}{\partial w} > 0; \tag{13}$$

$$\pi_{10}(w) + \pi_{01}(w) < 1. \tag{14}$$

Hausman, Abrevaya, and Scott-Morton (1998) refer to (14) as the monotonicity condition. To see the relevance of this condition note that:

$$E(X_i^w) = \Pr(X_i^w = 1) = \pi_{10}(w) + [1 - \pi_{01}(w) - \pi_{10}(w)]\pi_w. \tag{15}$$

Thus the monotonicity condition ensures that $\Pr(X_i^w)$ (reported use) increases in $\Pr(Y_i^w) = \pi_w$ (actual use).

The relationship between sub-period and target incidence is assumed known and for simplicity has the same form as used in the continuous case (i.e. $\pi_w = (w/S)p$). For discrete data this is a relatively crude representation for the imputation process that is best suited for situations where the population proportion is small when $w = 1$. Under these conditions, (2) remains the natural estimator of the target population proportion but to emphasize that it is now a proportion that is being estimated the sample statistics defined by (2) are denoted by $\hat{p}_w$.

With these assumptions, the following can be derived:

[1] In fact it may also not be appropriate in continuous case, see for example Bound, Brown and Mathiowetz, 2000.

(16) $$Bias(\hat{p}_w) = \pi_{10}(w)\left(\frac{S}{w} - p\right) - \pi_{01}(w)p$$

and

$$Var(\hat{p}_w) = Var\left[N^{-1}\left(\frac{S}{w}\right)\sum_{i=1}^{N} X_i^w\right] = \left(\frac{S}{w}\right)^2\left[\frac{\theta_w(1-\theta_w)}{N}\right]$$

so that

(17) $$MSE(\hat{p}_w) = \left[\pi_{10}(w)\left(\frac{S}{w} - p\right) - \pi_{01}(w)p\right]^2 + \left(\frac{S}{w}\right)^2\left[\frac{\theta_w(1-\theta_w)}{N}\right]$$

where $\theta_w = E(X_i^w)$.

As in the continuous case the discussion of results is organized around a number of cases.

Case 1: No recall bias

Here $\pi_{01}(w) = \pi_{10}(w) = 0$ and (17) simplifies to:

(18) $$MSE(\hat{p}_w) = \frac{p\left(\frac{S}{w} - p\right)}{N}$$

and it is clear that the optimal recall window is $w = S$. There is no gain from choosing a smaller recall window; such choices would involve collecting less data for the same sample size.

Case 2: Comparison of w =1 and w = S

Table 2 provides relative root mean square error (RMSE) measures for a particular set of parameters chosen to be: N =1000, S = 12, p = 0.6 and classification errors that range from zero to 0.3. Values in the table that are less than unity imply that the recall window of $w = S$ is preferred over $w = 1$ if the objective is estimation of the population rate of incidence over the target period.

There are clear patterns in these results that can be summarized as follows:

- Even in the presence of misclassifications errors, $w = S$ is preferred over $w = 1$ often strongly so, for a wide range of parameter configurations;
- From (16), notice that misclassifications may introduce bias that may be either positive or negative. Moreover, when classification errors are similar they tend to cancel each other out and such a situation tends to favour a longer recall window. This situation corresponds to entries on or near the main diagonal of the table.
- On the other hand $w = 1$ is preferred over $w = S$ when just one of the recall errors is large.
- Changes in bias also explains why, given a particular level of one type of misclassification error, there is no systematic movement in the preference for $w = S$ over $w = 1$ as the other misclassification rate is changed.

Table 2: Relative RMSE of alternative estimators of *p*: Case 2 with *N* = 1000, *S* = 12, *p* = 0.6*

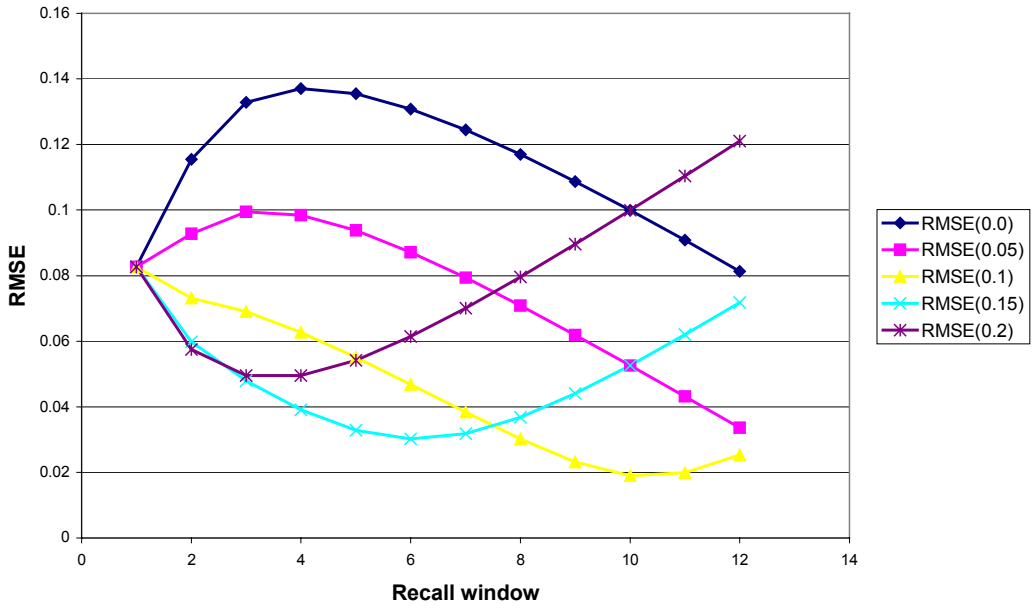
	$\pi_{10}(S)$						
$\pi_{01}(S)$	**0.00**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**
0.00	0.19	0.41	0.75	1.10	1.46	1.82	2.18
0.05	0.30	0.22	0.52	0.87	1.22	1.58	1.94
0.10	0.52	0.22	0.31	0.63	0.99	1.34	1.70
0.15	0.75	0.41	0.19	0.41	0.75	1.10	1.46
0.20	0.98	0.63	0.30	0.22	0.52	0.87	1.22
0.25	1.22	0.87	0.52	0.22	0.31	0.63	0.99
0.30	1.46	1.10	0.75	0.41	0.19	0.41	0.75

* Table entries represent $RMSE(\hat{p}_S) / RMSE(\hat{p}_1)$.

Case 3: Optimal w

Figure 2 depicts a range of $RMSE(\hat{p}_w)$ functions that show the types of patterns that can result from different parameter configurations. All of the reported functions assume $N = 1000$, $S = 12$, $p = 0.6$ with the different functions generated by allowing the proportion of each type of classification errors to vary but subject to the restriction that $\pi_{01}(S) + \pi_{10}(S) = 0.2$. Each RMSE function is indexed by $\pi_{01}(w)$ and these errors are assumed to vary linearly with w.

Figure 2: RMSE comparisons of probability estimates



Just as in the continuous case, these RMSE functions can follow a variety of shapes and typically are non-monotonic. Moreover, the optimal recall window can be $w = 1$ as it is for $\pi_{01}(S) = 0$, $\pi_{10}(S) = 0.2$; $w = S$ as it is for $\pi_{01}(S) = 0.05$, $\pi_{10}(S) = 0.15$; or the optimal window may lie between these two extremes as it is in the remaining configurations.

4. Some empirical results from Swedish data

To illustrate how the previously derived results can be used to inform survey design we provide an example that is based on information concerning recall error associated with the number of nights spent in hospital over the previous three months obtained from Statistics Sweden's Survey of Living Conditions (the ULF survey) which has been linked to the national Patient Register (the National Board of Health and Welfare). Every year, Statistics Sweden conducts systematic surveys of living conditions, in the form of one-hour personal interviews with randomly selected adults aged 16-84 years. Since 1975 around 7,000 individuals have been interviewed each year. The Patient Register includes information about ICD-codes, hospital admissions, and the total length of stay in hospital over the past three months.

Information on the length of stay in hospital (number of nights) over the last three months by 11,948 patients from the 1996 and 1997 surveys was compared with registry data to determine the accuracy of reporting over this period. For the purposes of this example we assume that the patient registry is the *gold standard* by which the accuracy of reported use can be judged. Based a comparison of these data: 11,368 people were classified as true negatives; 368 as true positives; 135 people as false positives; and 65 as false negatives. Of the 445 people that had at least one hospitalization, the mean (variance) in the number of nights over the last three months was 9.6 (172.6) based on patient registry data, and 8.8 (178.7) nights were reported in the survey.

To illustrate how these data can be used to derive an optimal recall length we assume that the maximum period of time over which there would perfect recall is one week (i.e. $\pi_{10}(1) = \pi_{01}(1) = 0$ and $h(1) = g(1) = 0$) and that the target period is three months (i.e 12 weeks). As there is no information on the degree of error for other periods linear interpolation was used to estimate errors for different values of w between the perfect recall and the target period.

Firstly, consider the discrete case where the statistic of interest is the proportion of people spending at least one night in hospital. Based on the registry data, 445 of the 11,948 surveyed (i.e $p = 0.037$) actually spent time in hospital. In terms of recall error, $\pi_{10}(12) = 135/11{,}503 = 0.01$ and $\pi_{01}(12) = 59/386 = 0.13$ and hence the proportion of people reporting a hospital stay over the last three months $E(X_i^w) = 0.044$. As the pattern of errors in the survey response is consistent with the monotonicity condition, (17) can be used to calculate the RMSE for different lengths of recall and these are shown in Figure 3(a). While the target period of 12 weeks has the lowest RMSE, suggesting that it is the optimal recall period, it should be noted the level of error is only 3% below using a one week recall period.

To illustrate the continuous case, consider estimating the mean number of nights in hospital for those individuals which the patient registry indicated had at least one hospital stay over the target period. Using the summary statistics for the registry and survey data, $h(w) = 9.6/8.8 - 1 = -0.09$ and $g(w) = 178.7/172.6 - 1 = 0.03$. Using (7) the RMSE for different lengths of recall is shown in figure 3(b). It is evident that

substantial reductions in the degree of error can be achieved by increasing the recall period beyond one week (i.e. perfect recall period) and that a recall period of around 8 weeks will minimize the RMSE. This would suggest that if one were concerned with estimating hospital use over the target period, a survey with a two month rather than a three month recall period may be optimal, although the gains in terms of reductions in RMSE are relatively modest.

Figure 3a: Discrete case: whether hospitalized in last three months

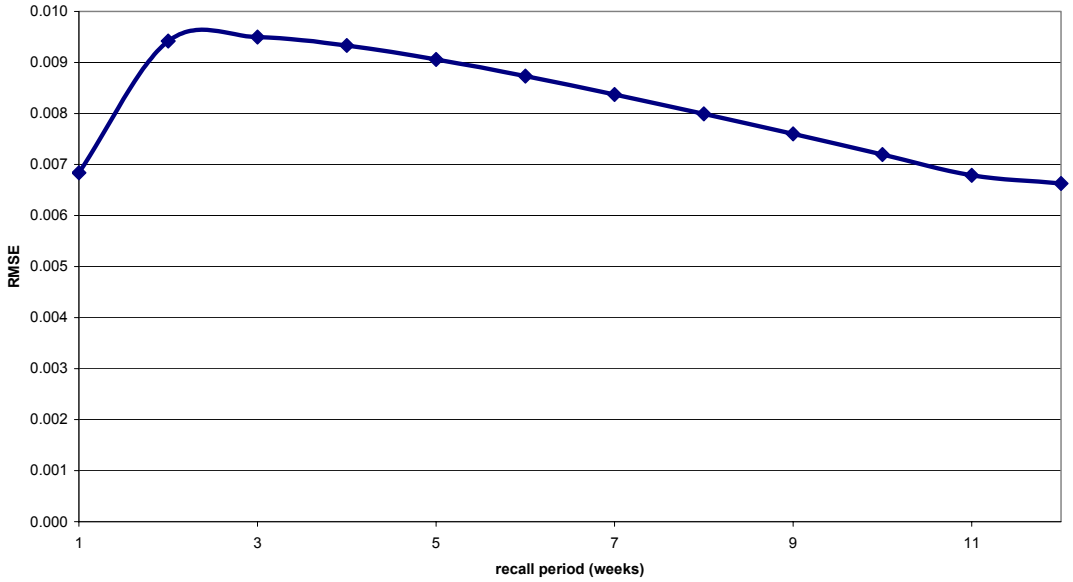
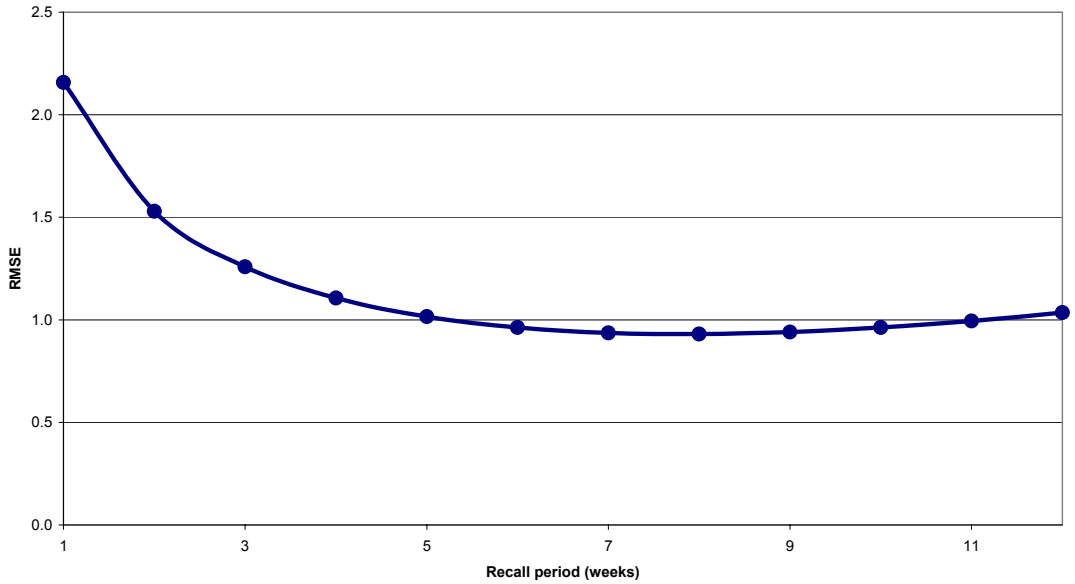


Figure 3b: Continuous case: No. Nights in hospital



5. Discussion

While stylised, the modelling serves to highlight the key aspects of the generic problem. There are tradeoffs. While longer recall windows may invariably introduce recall errors, using shorter recall windows also come with costs associated with less information. The analysis so far has served to illustrate these tradeoffs and to stress that the case for eliminating recall errors is not necessarily compelling.

It is worth highlighting some important dimensions of the problem that impact the analysis and results thus far. In particular, we have assumed a known imputation or prediction process for translating sub-sample data to an estimate of the key parameter for the target period window. For the discrete case there is a relationship between $\Pr(Y_i^w) = \pi_w$ and $\Pr(Y_i^S) = p$ and this was assumed known so that converting an estimate of π_w into an estimate of p did not introduce further error. In practice this is unlikely to be the case and having to infer say annual participation rates from monthly ones or having to somehow scale monthly consumption into a yearly estimate will surely introduce added estimation error. There may be cases where such errors are small, such as prescription drugs where consumption patterns are regular and hence very predictable. But there will be other situations where consumption is very irregular and/or subject to seasonal variation making prediction much more difficult.

Other things being equal, introducing an imputation process to deal with the missing data (Briggs et. al. 2003) is likely to favour collecting data over the target period as the default in survey design. Otherwise, the argument has to be that the researcher, with accurate sub-period data, can do a better job of predicting utilisation or consumption over the target period than the respondent does in recalling this information.

A key aspect of this line of argument is the distinction between the target period and a narrower period that eliminates recall error. Obviously if women can accurately recall whether they have had a Pap test in the last week and one week is a policy relevant and interesting period over which to measure Pap test incidence then the distinction is not overly important. However, in this example the policy relevant period is more likely to be one or two years rather than one or two weeks. It is also possible that

short windows designed to eliminate recall error may bring about other forms of error. For example, if one is concerned with estimating the number of nights in hospital then a short window has the potential to censor longer hospital stays and may introduce sample selection bias if the survey is undertaken in the community rather than while the patient is in hospital.

Thus far the emphasis has been on estimating a population mean or proportion. Clearly, there may be other objectives that could be considered in deciding on the optimal recall window. The data may be used in further analyses, say as the dependent variable or one of the explanatory variables in a regression analysis. Here we return to our original contention that econometricians devote considerable time and effort in developing procedures to counteract flawed data. Such efforts are usually in the context of data that has been collected by others and potentially for purposes other than econometric analysis; see Griliches (1986) for an extended discussion of this characterisation.

If researchers have the opportunity to be involved in the survey design then some of the problems often encountered with secondary data may be avoided. However, given procedures exist that can accommodate measurement errors of the type we have been discussing, their availability needs to be considered as part of the design problem. What does seem to be a sensible approach is to recognize the issues and be flexible in the design of questions and the collection of survey data in general. One such idea is the use of validation surveys. As Bound, Brown and Mathiowetz (2000, p. 3832) conclude:

> "… there are clear payoffs to greater involvement of users in the design of validation studies"

and that

> "… in general we believe that more effort devoted to collecting and analyzing validation data would significantly enhance the value of survey data."

There is no shortage of validation studies in health; for example see the studies surveyed in Bound, Brown and Mathiowetz (2000), Bowman, Sanson-Fisher and Redman (1997) and Evans and Crawford (1999). However, the emphasis is invariably

on determining the accuracy and validity of particular survey instruments and data collection modes. What is missing is the use of validation data in conjunction with large-scale surveys in order to improve inferences.

To reliably estimate appropriate recall periods in future health care surveys it is necessary to gain an understanding the nature of recall error for different types of health care. One way of estimating optimal recall periods would be undertake a randomised trial in which patients were asked to recall utilisation of different periods and these would then be compared with objective data on health care use, in order to estimate how the degree of error changes over time. Given variation in the frequency and pattern of use of health care (e.g. the proportion of the population visiting a physician versus attending a hospital over the previous year) and the potential for variation in the propensity to recall different types events it is likely that the optimal recall period may vary across different types of health care. Another aspect of survey design is to define the target period that is relevant for the research issue under consideration, as this will also influence the optimal recall period.

6. Conclusions

A common criticism of statistical analyses using self-reported data is the problem caused by recall error. A possible response by survey designers is to choose extremely short recall windows in order to eliminate this source of error. For example, some national health surveys only ask health service use over the previous two weeks. One possible interpretation of our results is that such a reaction may not be entirely justified. Recall error may be eliminated but at a potentially huge cost in terms of information loss. These tradeoffs need to be considered and factored into decisions on recall windows. Moreover, if recall errors in the form of recall bias are known to exist it is possible that alternative estimators may be employed to correct for the resultant biases. Such approaches were not included in our comparisons and if available would likely add support for using longer rather than shorter recall windows.

Similarly, we abstracted from the imputation problem that arises when short recall windows are used. If the ultimate objective is to measure consumption or utilization

over a target period that is much longer than the recall window then the data collected for the sub-period needs to be scaled up in some way in order to produce the desired estimate. This is likely to be a non-trivial exercise in many applications. Once again this consideration is likely to put more weight on choosing longer rather than shorter window lengths. In such cases, the trade-off is between the errors incurred by the respondent in recalling consumption over the target period compared with the error induced by the imputation process used by the researcher.

Relatively simple models have generated all calculations of optimal recall windows but the results are suggestive of a number of important relationships and tendencies. With some knowledge of the key parameters it may be possible to decide whether a shorter or longer recall window is better. But it is clear that there is no general answer to the question of optimal recall windows. It depends on the primary objective of the data collection and the full range of window lengths may be optimal for certain reasonable circumstances. What we have argued is that there is no compelling reason why the default situation in survey design should be a short recall period.

References

Bound, J., Brown, C. and Mathiowetz, N. (2000), "Measurement error in survey data" Ch. 59 of J.J. Heckman and E. Leamer eds. *Handbook of Econometrics*, Vol 5, North Holland, 2000.

Bowman, J.A., Sanson-Fisher, R. and Redman, S. (1997), "The accuracy of self-reprted Pap smear utilization", *Social Science and Medicine* 44, 969-976.

Benitez-Silva, H., Buchinsky, M., Chan, H.M., Cheidvasser, S. and Rust, J. (2004), "How large is the bias in self-reported disability?", *Journal of Applied Econometrics* 19, 649-670.

Briggs A.H., Clark T., Wolstenholem J., Clarke P. (2003), "Missing ... presumed at random: cost-analysis of incomplete data", *Health Economics*, 12, 377-392.

Carson, R., Groves, T. and Machina, M. (1999), "Incentive and Informational Properties of Preferences Questions," Plenary Address, European Association of Environmental and Resource Economists, Oslo Norway.

Deaton, A. (1997), *The analysis of household surveys: A microeconometric approach to development policy*, John Hopkins University Press.

Evans C. and Crawford B. (1999), "Patient self-report in pharmacoeconomic studies: Their use and impact on study validity", *Pharmacoeconomics*; 15(3): 241-256.

Griliches, Z. (1986), "Economic data issues", Ch. 25 of Z. Griliches and M.D. Intriligator eds. *Handbook of Econometrics*, North Holland.

Hausman, J.A., Abrevaya, J. and Scott-Morton, F.M. (1998), "Misclassification of dependent variable in discrete-choice setting", *Journal of Econometrics* 87, 239-270.

Health Equity Research Group, (2004), "Income-related inequality in the use of medical care in 21 OECD countries", Towards High Performing Health Systems: Policy Studies from the OECD Health Project, OECD, Paris.

Philipson, T. and Malani, A. (1999), "Measurement errors: A principal investigator-agent approach", *Journal of Econometrics* 91, 273-298.

Sudman, S. and Bradburn, N.M. (1973), "Effects of time and memory factors on response in surveys", *Journal of the American Statistical Association*, 68, 805-815.