School of Economics
UNSW, Sydney 2052
Australia

http://www.economics.unsw.edu.au

A Discrete Choice Model with Misclassification and Multiple Recall Periods

Rochelle Belkar, Waranya Pim Chanthapun and Denzil G. Fiebig

School of Economics Discussion Paper: 2007/10

The views expressed in this paper are those of the authors and do not necessarily reflect those of the School of Economic at UNSW.

978 0 7334 2445 8

ISSN 1323-8949

ISBN 978 0 7334 2445 8

A discrete choice model with misclassification and multiple recall periods*

Rochelle Belkar, Waranya Pim Chanthapun and Denzil G. Fiebig

School of Economics, University of New South Wales

Summary:

Self-reported data collected via surveys are often subject to measurement error caused by recall errors. While methods to minimize such problems are to be encouraged we argue that such errors are often unavoidable and need to be accommodated in estimation. Such methods have been developed to facilitate estimation in the case of a binary choice model with a misclassified dependent variable. These methods are extended to situations where the survey contains questions with multiple recall windows. The estimation procedure is illustrated using an analysis of Australian data on visits to a GP.

Keywords: recall error; errors in variables; binary choice; utilisation of GPs.

* The authors are grateful to Ryan Castle for valuable research assistance and to seminar participants at ANU, Monash University, UTS and University of South Australia for helpful comments. We would also like to thank two anonymous referees for constructive suggestions that have led to improvements in the paper. This research was partially supported by the NHMRC through a Program Grant No. 254202.

Version: 13 February, 2007

Corresponding author:
Professor Denzil G Fiebig
School of Economics
University of New South Wales
Sydney, 2052, Australia
FAX: +61-2-9313 7691 email: d.fiebig@unsw.edu.au

Introduction

Measurement error is a pervasive problem in the collection of survey data. Recall error, where respondents provide inaccurate answers when recalling past events, is a major source of such problems; see for example Sudman and Bradburn (1973). Consider the case of the utilisation of various health services such as whether individuals have recently visited a doctor or had a Pap test. In such cases, survey designers aiming to reduce or eliminate recall error have an incentive to choose very short recall windows and ask whether a respondent has visited a GP in say the last 2 weeks or not.

While respondents are likely to be able to better recall visits over the last two weeks compared with say the last 6 months, Clarke, Fiebig and Gerdtham (2005) argue that the reduction in recall error will have an associated cost in terms of information loss. Moreover, they note that often a recall window that minimizes recall error will not be very policy relevant. For instance, women may be able to accurately recall whether they have had a Pap test in the last month but the policy relevant period is whether they have been tested in the last two years.

Self-reported data collected via surveys are often subject to measurement error and while methods to minimize such problems are to be encouraged we contend that such errors are often unavoidable. As such procedures to cope with measurement error problems need to be developed and utilized; see for example the excellent survey by Bound, Brown and Mathiowetz (2000). When the utilisation variable is dichotomous and is the endogenous variable, Hausman, Abrevaya and Scott-Morton (1998) developed maximum likelihood (ML) and semi-parametric estimation procedures for a binary choice model where there is some probability that the choices are misclassified. These methods have been used by Caudill and Mixon (2005), Kenkel, Lillard and Mathios (2004) and by Leece (2000) and extended to ordered choice problems by Dustman and van Soest (2004). See also Hsiao and Sun (1999).

Our primary aim is to develop and illustrate an extension of the Hausman, Abrevaya and Scott-Morton (1998) ML estimation procedure when the survey contains utilisation questions over multiple recall windows. For example, respondents are asked if they have visited a GP in the last two weeks and if not whether they have been in the last six months or not. Responses to the second question are likely to be subject to recall error and the procedures developed by Hausman, Abrevaya and Scott-Morton (1998) appropriate. However, it is reasonable to assume that responses to the first utilisation question are error-free and its availability provides accurate information (for a subset of observations) that is exploited in our suggested estimation approach. Interpreted in this way, the approach suggests an alternative to the collection of validation data and thus, for those involved in survey design, represents a very simple and cost effective means of validating part of the data.

A survey designer would ask two utilisation questions over different recall windows. First, they put aside issues of recall error and decide on the most relevant and useful utilisation period for analysis. For example, there may be policy relevant periods such as recommended screening intervals. Then the narrow recall window would be chosen in order to generate error-free measures.

In what follows, we develop the estimation procedure and illustrate and evaluate its use in an analysis of Australian data on visits to a GP.

The econometric model

Consider a binary variable y_i that represents utilisation over some policy relevant target period and is generated according to:

$$(1)\quad \Pr(y_i = 1 \mid x_i) = F(x_i'\beta) = p_i$$

for i=1, … , n respondents and where x_i represents a vector of explanatory variables and $F(.)$ is a known (usually normal or logistic) cdf. Assume y_i is error-free but is not observed. Instead two binary variables y_{1i} and y_{2i} are observed where the latter is an error-ridden version of y_i with misclassification errors. In particular,

$$(2)\quad y_{2i} = d_i y_i + (1 - d_i)(1 - y_i)$$

where d_i is an unobserved binary indicator of whether the ith response is correctly classified or not. The second binary variable, y_{1i} is assumed to be an error-free measure of utilisation but over a recall period that is shorter than that associated with y_i and y_{2i}. Because of the nesting structure of the recall windows, y_{1i} = 1 implies y_{2i} = y_i = 1 and thus y_{1i} identifies true positives amongst the error-ridden y_{2i} observations. A subset of observations where d_i = 1 is observed and it is this extra information that we seek to exploit. It would be possible to extend the current analysis to a situation where the second binary variable, y_{1i} is not error-free but where misclassification errors are less likely than in the longer recall period. Given our emphasis on survey design where the narrow recall window would be chosen to avoid this case, we leave this extension for future work.

The model specification is completed by assuming a relationship between the observed, short-recall measure, y_{1i}, and the unobserved error-free measure, y_i. Two possible specifications seem natural:

$$(3') \quad \Pr(y_{1i} = 1 \mid x_i) = \alpha p_i; \quad 0 \le \alpha \le 1$$

or

$$(3'') \quad \Pr(y_{1i} = 1 \mid x_i) = F(-\mu + x_i'\beta); \quad \mu > 0.$$

When no misclassification is assumed (3ʹ) implies a multinomial regression model with three outcomes: (y_{1i} =1, y_{2i} =1), (y_{1i} =0, y_{2i} =1) and (y_{1i} =0, y_{2i} =0). In the particular case where $F(.)$ is assumed to be logistic, this special case corresponds to a multinomial logistic (MNL) model with pooled states; see Cramer and Ridder (1991). Alternatively, (3″) represents an ordered regression model with the outcomes representing three levels of duration since last utilization. The choice between these two alternatives is essentially an empirical matter and for the application to be provided below, the second specification in (3″) is not supported by the data. Thus, in what follows it is (3ʹ) that is developed to accommodate misclassification errors. This may not always be the case so we are not advocating that this should be the only possible specification considered to complete the model when using multiple recall windows as a means of internal validation of possibly error-ridden data.

The misclassification probabilities are defined as:

$$
\begin{aligned}
(4)\quad \pi_{10} &= \Pr(d_i = 0 \mid y_i = 0), \quad \text{and} \\
\pi^*_{01} &= \Pr(d_i = 0 \mid y_i = 1) \\
&= \Pr(y_{1i} = 0 \mid y_i = 1)\Pr(y_{2i} = 0 \mid y_i = 1, y_{1i} = 0) \\
&= (1-\alpha)\pi_{01}
\end{aligned}
$$

where α represents the proportion of successes (y_i = 1) occurring in the narrow recall period, π_{10} is the probability of a false positive and π^*_{01} is the (unconditional) probability of a false negative. When y_{1i} is observed, define π_{01} as the probability of a false negative conditional on y_{1i} = 0. Note that the classical errors-in-variable framework associated with continuous random variables is not appropriate in the discrete case being considered here. The measurement errors represented by (4) are clearly not independent of the true variable and in fact will be negatively correlated with y.

Given these assumptions the implied generating process for y_{2i} is given by

$$(5)\quad \Pr(y_{2i} = 1 \mid x_i) = \pi_{10} + [1-(1-\alpha)\pi_{01} - \pi_{10}]p_i.$$

α = 0 corresponds to the case of no y_1 data and the analysis relies solely on the mismeasured y_2 data. This is the situation considered by Hausman, Abrevaya, and Scott-Morton (1998).

With data on y_{1i} and y_{2i}, the three dummy variables representing the possible outcomes are:

$w_{11i} = y_{1i}\, y_{2i}$, w_{01i} = (1-y_{1i})y_{2i} and w_{00i} =(1-y_{1i})(1-y_{2i}).

Then it follows

$$
\begin{aligned}
(6)\quad \Pr(w_{11i} = 1 \mid x_i) &= \alpha p_i = p_{11i} \\
\Pr(w_{01i} = 1 \mid x_i) &= (1-\alpha)p_i(1-\pi_{01}) + (1-p_i)\pi_{10} = p_{01i} \\
\Pr(w_{00i} = 1 \mid x_i) &= (1-\alpha)p_i\pi_{01} + (1-p_i)(1-\pi_{10}) = p_{00i}
\end{aligned}
$$

and the log-likelihood function is given by:

$$(7) \quad \log L(\alpha,\beta,\pi_{10},\pi_{01}) = \sum_{i=1}^{n}\left(w_{00i}\log p_{00i} + w_{01i}\log p_{01i} + w_{11i}\log p_{11i}\right).$$

ML estimation using (7) will not be able to distinguish between two sets of parameter values; (α, β, π^*_{01}, π_{10}) and (α, $-\beta$, $1-\pi^*_{01}$, $1-\pi_{10}$). Thus for identification, we also assume that $\pi^*_{01} + \pi_{10} < 1$. Hausman, Abrevaya, and Scott-Morton (1998) refer to this as the monotonicity condition, which from (5), we see ensures that reported use increases in actual use.

Application to modelling utilisation of GP services

The endogenous variables are whether or not a respondent has visited a GP in the last 2 weeks (y_1= *GP1*) or in the last six months (y_2= *GP2*). We expect utilisation to be affected by personal characteristics including age, household income, gender, their general health status, education, location, ethnicity, and whether they have private health insurance. Table 1 provides a brief description of the actual variables to be used in the analyses together with their sample means.

These data are taken from the 2001 National Health Survey (NHS) which is the fifth health survey of its type conducted by the Australian Bureau of Statistics (ABS). The 2001 NHS was conducted using a sample of 17,918 private dwellings across Australia representing a response rate of 92%. Within each sampled dwelling a random selection of usual residents were chosen including one adult, all children aged 6 or less and one child aged 7-17. In our analysis all children were excluded leaving a sample of 17,918 (adult) observations available for estimation. In a small number of cases (just over 1%) adult respondents were unable to answer for themselves and the person responsible for them was used as a proxy to answer questions. The data set contains 3,355 missing observations on the household income question accounting for approximately 19 percent of the sample. These 'item non-response' observations are retained but are flagged by the introduction of the dummy variable 'HINCMISS' into the covariate set.

We first estimate a binary logit model for *GP2* assuming no misclassification errors. These results are given in the first column of Table 2. In order to gain further insights into the magnitude of some of these effects, the results have been translated into impacts on probabilities. The estimated probability for a baseline case is determined. Then, characteristics are varied one at a time, and the new probability estimate is calculated. The probability estimates for three sets of results are reported in Table 3. Again, at this stage we only consider the relevant results provided in the column labelled "logit".

As would be expected under Australia's Medicare system, health related factors, namely age and self-assessed health status (SAHS), are the important determinants of GP utilisation. Although income is not found to be significant, other socioeconomic and demographic factors such as gender, ethnicity, education and area of residence are found to have a significant impact on GP utilisation. The baseline case in the probability calculations in Table 3 refers to a 40 year old male, in the lowest income decile who has fair SAHS, less than complete high school, has no private health insurance, is not Australian born and lives in a metropolitan area. The logit estimate for the probability that such an individual visited a GP in the last 6 months is estimated to be 0.761 with the largest change occurring when the SAHS is changed from fair to excellent in which case the estimated probability drops to 0.396. The age and SAHS profiles and the other predicted probabilities are sensible in terms of expected changes in probabilities relative to the base case but the magnitudes are potentially biased if misclassification errors are present.

The Hosmer-Lemeshow (1989) goodness-of-fit test is performed with observations ordered into 20 groups by their expected probability of observing $y_2 = 1$. The resulting test statistic of 32.50 is above the relevant 5% critical value for a chi-square distribution with 18 degrees of freedom, which is 28.9. Thus, the GP utilisation binary logit model with no misclassification is rejected at the 5% level. Figure 1 shows the expected and observed relative frequencies of GP utilisation in 20 classes of equal observations. The graph shows a deviation of the expected frequencies from the actual ones at the low end of the expected probability. This could be interpreted as evidence of misclassification, in particular the presence of false positives whereby respondents who have not been to a GP in the last six months say they have.

Now consider the possibility of misclassification in the variables representing visits in the last six months. The second column of Table 2 provides (EIV logit) estimates for the binary logit model with possibilities of misclassification developed by Hausman, Abrevaya and Scott-Morton (1998). The result is consistent with what we observe from Figure 1. The probability of false positives is estimated to be 0.33 whereas the probability of false negatives is small and statistically insignificant. The systematic difference in coefficient estimates in the first two columns of Table 2 and the difference in probability estimates in the first two columns of Table 3 is attributable to biases caused by not accounting for misclassification errors. For example, the predicted probabilities generated by the EIV logit for a GP visit in the last 6 months for the base case is 19% lower than that predicted by the logit model. The percentage differences are even more pronounced for rarer events. When the base case is changed to be in excellent health the predicted probability of EIV logit is 0.159 which is 60% lower than that predicted by the logit model.

To illustrate the extension of the Hausman, Abrevaya and Scott-Morton (1998) ML estimation procedure when the survey contains utilisation questions over multiple recall windows, we follow the procedure outlined above. In particular, we exploit the availability of presumably recall error-free information on whether or not a respondent has visited a GP in the last two weeks. Support for this assumption is provided by application of the Hosmer-Lemeshow test to a binary logit model for *GP1* where the test statistic is found to be 16.18 with a *p*-value of 0.58, which is consistent with no mis-specification. The last column of Table 2 provides (MRW logit) estimates for the model presented in (7).

Compared to the coefficient estimates for the binary logit with single recall window and misclassification, the estimates obtained from the multiple-recall-window model display the same signs, consistent statistical significance and similar magnitudes. Thus, in Table 3 it is not surprising to see that they also produce very similar estimated probabilities. However, the multiple-recall-window model gives more precise estimates as indicated by uniformly smaller standard errors of the coefficient estimates. On average the MRW standard errors are approximately 15% less than those for EIV logit. This is an efficiency gain that results from exploiting extra information on the utilisation over a shorter recall period. The information provided by the *GP1* responses effectively reduces the number of *GP2*=1 responses potentially misclassified. This information also translates into a much more precise estimate of the probability of a false positive which, with an estimated value of 0.40 is again found to be substantial. Just as for the EIV logit, the MRW logit produces an estimate of the probability of false negatives that is small and statistically insignificant.

The additional parameter estimated with the MRW logit model is α. If there were no errors of misclassification then the MLE of α is simply the ratio of the sample proportions for y_1 and y_2 which for our data yields 0.28/0.73 = 0.38. Thus of those reporting at least one GP visit in the last six months, 38% reported a visit in the last 2 weeks. Because a propensity for false positives has been found, this is an underestimate of this proportion and the MRW estimation results yield an estimate for α of 0.50.

As mentioned in the development of the MRW model, (3″) represents an alternative specification of the relationship between the observed, short-recall measure, y_{1i}, and the unobserved error-free measure, y_i. Under this specification the binary logit using y_{1i} alone is able to provide consistent coefficient estimates. This is the "parallel regressions" assumption of ordered regression models; see for example Long (1997). Using only y_{1i} presents a problem with estimating the intercept as it will not be possible to separately identify the additional parameter μ and the intercept in the longer recall model; only their sum will be estimable. This is a problem if the aim is to predict utilisation probabilities but not if estimates of marginal effects are all that is required.

For our current purposes the parallel regressions assumption is the basis for a convenient diagnostic check. If (3″) did in fact provide a good approximation for our data then we would expect that coefficient estimates generated by a binary logit using the narrow recall variable, *GP1*, should yield coefficient estimates that are approximately equal to those estimated with using *GP2*. On the other hand our current specification given in (3′) implies that the estimates from binary logit using *GP1*, should be systematically less in magnitude than the EIV logit estimates and this is in fact what we find; results for the former are not reported but are available on request. The ratios of binary logit estimates for *GP1* and EIV logit estimates for all 25 coefficients (except the intercept) have a mean equal to 1.04 but a median of only 0.63 because the mean is inflated by several outliers associated with coefficients that are not precisely estimated. Taking just the coefficients associated with the 16 EIV logit estimates in Table 2 that are significant at the 5% level, all have ratios less that unity with a mean of 0.49 and median of 0.46. For our particular application the alternative specification using (3″) is not supported and so was not estimated.

Simulation evidence

In order to provide further evidence on the potential usefulness of the proposed approach we use our application as the basis of a small simulation study. Our application has many estimated coefficients making it difficult to use directly as the basis for a simulation study. To simplify matters a single variable is produced corresponding to the estimated value of the index in equation (1). Thus, the true data generating process for the unobserved error-free variable is assumed to be:

$$(8) \quad \Pr(y_i = 1 \mid z_i) = F(z_i\delta) = p_i,$$

where $z_i = x_i'\hat{\beta}$ with the estimated coefficients coming from the multiple recall window results of Table 2. δ is taken to be unity in all experiments and while an intercept is estimated in all models the true value is zero when generating the simulated choice data.

Probabilities of misclassification are varied as part of the experimental design but the overall amount of misclassification remains fixed at the level found in our application so that $\pi_{10} + \pi_{01} = 0.4$. α, which governs the amount of internal validation made available by the narrow recall window, takes on values of 0.35 and 0.5. Given a particular design point we generate 100 Monte Carlo replications of y_1 and y_2 and then use y_2 to estimate a logit model with and without accommodating for misclassification and use both y_1 and y_2 to estimate the multiple recall window logit. The sample size is fixed at n = 17,918 for all distinct design points.

The three estimators of δ are compared on the basis of three criteria: the finite sample bias, relative efficiency as measured by the root mean squared error (RMSE) and a measure of how well the asymptotic standard errors (ASE) reflect the true variability of the estimators, represented by the ratio of the ASE and the RMSE. These results are presented in Table 4. There were some problems with convergence of the EIV and MRW estimation procedures and these cases were not included in the analysis. As a result some of the reported statistics are based on less than 100 observations. The problem was more prevalent for the EIV algorithm and thus one benefit of the extra information utilized with MRW is improvement in convergence. Across all design points and all iterations, the EIV algorithm failed to converge on 46 (7.6%) occasions while the MRW algorithm failed only 4 (0.7%) times. The vast majority of convergence failures occurred when π_{10} = 0.1 which corresponds to design points when there is likely to be fewer observations in the tails (predicted probabilities close to zero or unity) making it more difficult to estimate smaller amounts of misclassification.

As expected there is clear downward bias in the logit estimates. By contrast there is little bias in either the EIV or MRW logit estimates with all means over the replications close to the true value of unity. Where there is substantial improvement in MRW with respect to EIV is in terms of relative efficiency. The RMSE's of MRW logit are systematically below those of EIV logit. Because of the substantial biases, the logit is clearly dominated by both EIV and MRW logit. The asymptotic standard errors of EIV and MRW logit do a reasonable job of reflecting the true variability of the respective estimators. However, EIV logit tends to produce ASE's that are larger than the true variability and thus tend to be too conservative. MRW logit is somewhat better performed on this metric, although does sometimes have a tendency to understate true variability.

Admittedly this represents a limited exploration of the finite sample properties of these estimators and a more extensive investigation would be required to provide more refined recommendations. However, some broad conclusions are supported by the evidence. The MRW approach does seem to provide gains over the EIV approach in terms of computational robustness and relative efficiency. These key features of the results are consistent over all of the design points.

Conclusion

This study has extended the Hausman, Abrevaya and Scott-Morton (1998) ML estimation procedure when the survey contains questions over multiple recall windows in an effort to better accommodate measurement error caused by recall errors. The estimation procedure developed is illustrated in an analysis of Australian data on GP utilisation and by way of a small simulation study. We found that exploiting information on GP utilisation over multiple recall windows in the estimation procedure improves the precision of the estimates and this is supported by the Monte Carlo simulations.

If self-reported survey responses could be routinely validated using say medical provider records then problems caused by measurement errors could be eliminated or at least minimized. However, the combination of comprehensive survey data and accurate utilisation data from external sources is not common. For those involved in survey design, the use of multiple recall windows may be a simple and useful device to provide more precise estimates when recall errors are likely to occur in asking utilisation questions over policy-relevant recall periods and where validation samples are not available.

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.

Caudill, S.B. and Mixon, F.G. (2005), “Analysing misleading discrete responses: A logit model based on misclassified data”, *Oxford Bulletin of Economics and Statistics* 67, 105-113.

Clarke, P.M., Fiebig, D.G. and Gerdtham, U.G. (2005), “Optimal recall length in survey design”, *CAER working paper, 2005/6*, University of New South Wales.

Cramer, J.S. and Ridder, G. (1991), “Pooling states in the multinomial logit model”, *Journal of Econometrics* 47, 267-272.

Dustman, C. and van Soest, A. (2004), “An analysis of speaking fluency of immigrants using ordered response models with classification errors”, *Journal of Business and Economic Statistics* 22, 312-321.

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.

Hosmer, D. W. and Lemeshow, S. (1989), *Applied Logistic Regression,* New York: John Wiley and Sons.

Hsiao, C. and Sun, B.-H. (1999), “Modeling survey response bias – with an analysis of the demand for an advanced electronic device”, *Journal of Econometrics* 89, 15-40.

Kenkel, D.S., Lillard, D.R. and Mathios, A.D. (2004), “Accounting for misclassification error in retrospective smoking data”, *Health Economics* 13, 1031-1044.

Leece, D. (2000), “Household choice of fixed versus floating rate debt: A binomial probit model with classification error”, *Oxford Bulletin of Economics and Statistics* 62, 61-82.

Long, J.S. (1997), *Regression Models for Categorical and limited Dependent Variables*, Sage Publications, Inc.

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

Table 1: Data description and means

Variables	Description	Mean
GP1	1 if consulted a GP in the last two weeks	0.28
GP2	1 if consulted a GP in the last six months	0.73
AGEYRS	Age in years	46.78
HINC1	1 if household income in first decile	0.12
HINC2	1 if household income in second decile	0.09
HINC3	1 if household income in third decile	0.09
HINC4	1 if household income in fourth decile	0.08
HINC5	1 if household income in fifth decile	0.08
HINC6	1 if household income in sixth decile	0.07
HINC7	1 if household income in seventh decile	0.07
HINC8	1 if household income in eighth decile	0.07
HINC9	1 if household income in ninth decile	0.07
HINC10	1 if household income in tenth decile	0.07
HINCMISS	1 if household income figure missed	0.19
FEMALE	1 if female	0.54
HLTHEX	1 if self assessed health status excellent	0.17
HLTHVG	1 if self assessed health status very good	0.32
HLTHG	1 if self assessed health status good	0.31
HLTHF	1 if self assessed health status fair	0.15
HLTHP	1 if self assessed health status poor	0.05
TERT	1 if tertiary qualifications	0.16
DIPLOMA	1 if diploma	0.09
TRADE	1 if trade	0.27
NOQUAL	1 if only high school qualification	0.48
METRO	1 if reside in the metropolitan regions	0.66
INNER	1 if reside in the inner regions	0.21
OTHERREG	1 if reside in other regions	0.13
AUBORN	1 if Australian born	0.73
HPHI	1 if have hospital private health insurance	0.47
APHI	1 if have ancillary private health insurance	0.40

Table 2: GP Utilisation: Logit Estimation Results

	Binary logit GP2		Multiple recall windows with misclassification (MRW logit)
	Without misclassification (logit)	With misclassification (EIV logit)	
Constant	3.3956** (0.2247)	3.0806** (0.3308)	3.0061** (0.3078)
AGEYRS	-0.0832** (0.0074)	-0.1003** (0.0131)	-0.0889** (0.0099)
AGEYRS^2	0.0011** (0.0001)	0.0014** (0.0002)	0.0012** (0.0001)
HINC2	0.1830* (0.0963)	0.2502** (0.1241)	0.2452** (0.1147)
HINC3	0.0932 (0.0917)	0.1621 (0.1195)	0.1129 (0.1113)
HINC4	-0.1374 (0.0889)	-0.1435 (0.1208)	-0.1666 (0.1124)
HINC5	-0.0798 (0.0894)	-0.1116 (0.1201)	-0.2009* (0.1128)
HINC6	-0.2403** (0.0899)	-0.3644** (0.1268)	-0.3473** (0.1129)
HINC7	-0.1088 (0.0918)	-0.1437 (0.1254)	-0.2375** (0.1176)
HINC8	-0.0009 (0.0938)	-0.0305 (0.1253)	-0.1293 (0.1182)
HINC9	-0.0278 (0.0949)	-0.0677 (0.1273)	-0.1563 (0.1201)
HINC10	-0.1480 (0.0966)	-0.1945 (0.1333)	-0.2804** (0.1251)
HINCMISS	-0.2130** (0.0764)	-0.3002** (0.1032)	-0.3782** (0.0943)
FEMALE	0.7580** (0.0366)	1.0376** (0.1070)	0.9665** (0.0514)
HLTHEX	-2.2919** (0.1500)	-2.9318** (0.3068)	-3.0549** (0.1874)
HLTHVG	-1.8617** (0.1479)	-2.2431** (0.2482)	-2.5166** (0.1802)
HLTHG	-1.4748** (0.1479)	-1.7208** (0.2188)	-1.9907** (0.1758)
HLTHF	-0.7104** (0.1568)	-0.7941** (0.1928)	-0.9485** (0.1725)
TERT	0.0863 (0.0537)	0.1164 (0.0752)	-0.0064 (0.0713)
DIPLOMA	0.2292** (0.0657)	0.3285** (0.0940)	0.2755** (0.0842)
TRADE	0.1350** (0.0444)	0.2006** (0.0642)	0.1867** (0.0574)
INNER	-0.2280** (0.0463)	-0.3004** (0.0705)	-0.2892** (0.0609)
OTHERREG	-0.2780** (0.0543)	-0.4042** (0.0858)	-0.4761** (0.0731)
AUBORN	0.1244** (0.0423)	0.1952** (0.0614)	0.1347** (0.0542)
HPHI	0.1033* (0.0569)	0.1653** (0.0811)	0.1184 (0.0751)
APHI	0.0605 (0.0556)	0.0798 (0.0776)	0.1147 (0.0733)

α			0.4982** (0.0091)
π_{01}		0.0058 (0.0083)	0.0000 (0.0125)
π_{10}		0.3349** (0.0481)	0.4033** (0.0102)
Log-likelihood	-9236.24	-9228.80	-17701.21

The asterisks (*) and (**) indicate significant at 0.10 and 0.05 levels, respectively. Because the probability parameters are bounded, standard t-tests are not strictly appropriate but could be legitimately taken to be testing whether the probabilities are arbitrarily small.

Table 3: Predicted probability of visiting a GP in the last six months for different types of people using alternative estimation methods

Variable	logit	EIV logit	MRW logit
Baseline case	0.761	0.616	0.619
Variation in age			
Age = 20 years	0.813	0.696	0.685
Age = 30 years	0.769	0.626	0.624
Age = 40 years	0.761	0.616	0.619
Age = 50 years	0.793	0.669	0.672
Age = 60 years	0.852	0.771	0.767
Age = 70 years	0.916	0.880	0.872
Variation in self assessed health status			
Excellent	0.396	0.159	0.165
Very good	0.502	0.273	0.253
Good	0.598	0.388	0.365
Fair	0.761	0.616	0.619
Poor	0.867	0.780	0.808
Variation in income			
Household income in tenth decile	0.734	0.569	0.552
Variation in gender			
Female	0.872	0.819	0.811
Variation in birthplace			
Australian born	0.853	0.703	0.701

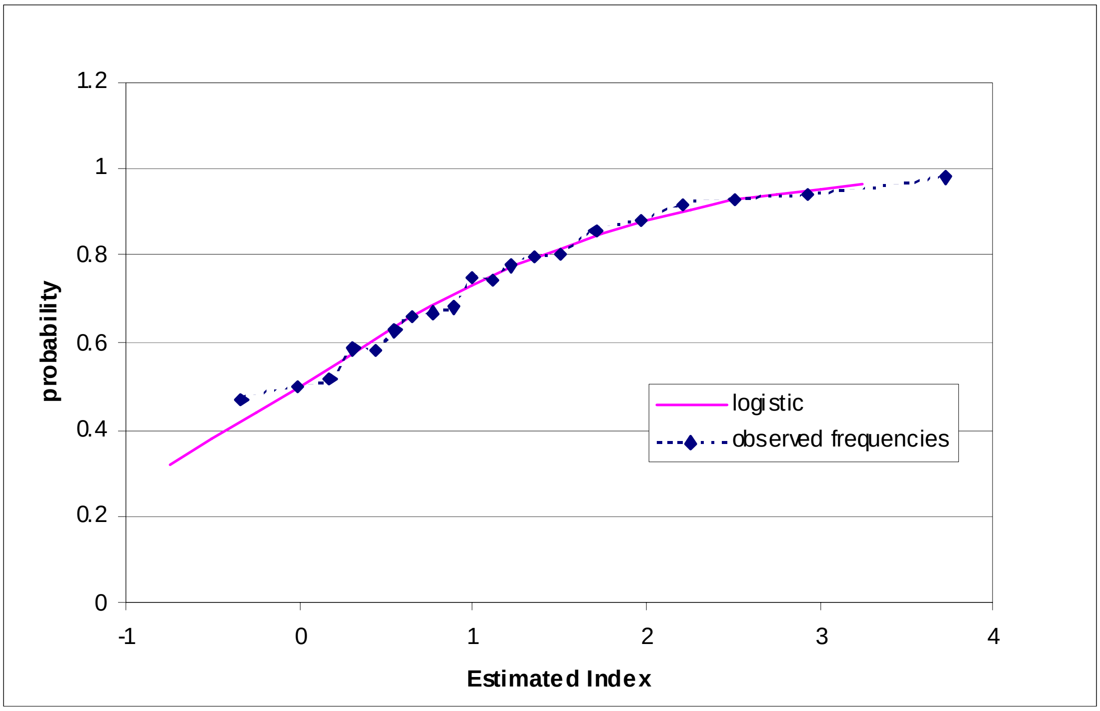
Notes:

- The baseline case is a 40 year old male, in the lowest income decile who has fair SAHS, less than complete high school, has no private health insurance, is not Australian born and lives in a metropolitan area.
- Only selected variations have been tabulated and the baseline case has been repeated when varying age and SAHS in order to better see the gradients.

Table 4: Monte Carlo simulation results

(π_{10}, π_{01})	$\alpha = 0.35$			$\alpha = 0.5$		
	logit	EIV logit	MRW logit	logit	EIV logit	MRW logit
			Means			
(0.1, 0.3)	0.461	1.023	1.003	0.624	1.033	1.002
(0.2, 0.2)	0.425	1.008	1.001	0.616	1.003	0.996
(0.3, 0.1)	0.401	1.013	1.000	0.645	1.018	1.003
			RMSE			
(0.1, 0.3)	0.539	0.126	0.054	0.376	0.095	0.049
(0.2, 0.2)	0.575	0.137	0.054	0.384	0.105	0.043
(0.3, 0.1)	0.599	0.178	0.060	0.356	0.085	0.041
			ASE/RMSE			
(0.1, 0.3)	0.024	1.066	1.037	0.038	1.054	0.874
(0.2, 0.2)	0.022	1.084	1.040	0.038	0.980	0.975
(0.3, 0.1)	0.022	0.922	0.962	0.045	1.181	0.975

Figure 1: Expected and observed frequencies of GP visits in 20 classes


probability
Estimated Index
logistic
observed frequencies