An Efficient Estimator for Dealing with Missing Data on Explanatory Variables in a Probit Choice Model[*]

Denis Conniffe and Donal O'Neill[**]
Department of Economics, National University of Ireland Maynooth

Abstract. A common approach to dealing with missing data in econometrics is to estimate the model on the common subset of data, by necessity throwing away potentially useful data. In this paper we consider a particular pattern of missing data on explanatory variables that often occurs in practice and develop a new efficient estimator for models where the dependent variable is binary. We derive exact formulae for the estimator and its asymptotic variance. Simulation results show that our estimator performs well when compared to popular alternatives, such as complete case analysis and multiple imputation. We then use our estimator to examine the portfolio allocation decision of Italian households using the Survey of Household Income and Wealth carried out by the Bank of Italy.

JEL Classification: C25, G11
Key Words: Missing Data, Probit Model, Portfolio Allocation, Risk Aversion.

[*] We would like to thank Olive Sweetman and seminar participants at the NUI Maynooth Economics seminar series and the 2008 Irish Economics Association annual conference for helpful comments on an earlier draft of this paper.
[**] Corresponding author: e-mail: donal.oneill@nuim.ie, tel:353-1-7083555.

1. Introduction

Many publications advocating approaches for dealing with missing data in regression type analyses have appeared in both the econometrics and mainstream statistical literature. Reviews of the latter are contained in Little (1993), Schafer (1997), Allison (2001) and Little and Rubin (2002). In the econometrics literature, relevant papers commence from Dagenais (1973), continuing through Gourieroux and Monfort (1981) and Conniffe (1983), and more recently Horowitz and Manski (2006), with a recent overview provided in Cameron and Trivedi (2005). Yet enthusiasm for the practical application of the methods seems muted at best. To quote the popular textbook by Wooldridge (2006), page 326:

There are ways to use the information on observations when only some variables are missing, but this is not often done in practice. The improvement in the estimators is usually slight, while the methods are somewhat complicated. In most cases, we just ignore the observations that have missing information.

While there are instances where this may be true, particularly when the proportion of incomplete data is small, there are many circumstances when it is unlikely to be the case. A well-known case arises when the regressors in the model are orthogonal. Consider a situation where the dependent variable, *Y*, and an explanatory variable, *x*, are recorded for the full sample of *n* observations but another explanatory variable, *w*, is only recorded for a subset, *r*, of the original sample. When *w* and *x* are orthogonal, we know that the simple regression of *Y* on *x* over all n observations, ignoring *w*, is the appropriate minimum variance estimator of the effect of *x* on *Y* conditional on *w*. A regression of *Y* on *x* and *w* for the complete cases would result in a similar point estimate but with a higher variance. Clearly if *n* is large relative to *r* the gain from employing the extra $(n - r)$ observations could be very substantial,

with the ratio of variances asymptotically of order $\frac{r}{n}$. This argument obviously generalises to multiple *x* and multiple *w* variables.

In practice, *w* and *x* are unlikely to be orthogonal, but it seems reasonable that if we were to assume that the *r* and (*n*– *r*) observations could be regarded as random samples from the same population we may be able to combine information available from *both* the full and complete case samples so as to obtain more precise estimates. The appropriate implementation formulae for the linear regression case have been presented in the papers cited earlier and the potential for precision improvement demonstrated there. In this paper we take a likelihood based approach that gives efficient estimators even when the *Y* variable is unobserved except for its sign. The approach also reproduces existing results for other models with this missing value problem, including linear regression, but the paper concentrates on the probit model. We provide straightforward, explicit, formulae for efficient coefficient estimators and their variances, which have not appeared previously in the literature for the probit model. We show, both by simulation and by analysis of real data, that our estimator outperforms alternative approaches, such as complete case analysis and multiple imputation techniques, for the given data structure. Our approach, with its explicit formulae for estimators and variances, also has virtues of transparency.

As with all approaches to dealing with missing data our estimator requires assumptions concerning the randomness of missing values. There is a large literature of considerable antiquity dealing with types of missingness. For example, Rubin (1974, 1976) outlines much of the basic terminology that has since been adopted and discusses the consequences of alternative patterns of missingness. In keeping with the majority of existing approaches we assume that the data are missing at random (MAR).[1] Data on *w* are said to be missing at random if the probability of missing data on *w* is unrelated to the value of *w*

[1] Horowitz and Manski (2006) discuss the construction of parameter bounds in the worst case scenario where the researcher has no prior information about the parameter of interest or the process that generates the missing data. In this conservative case small increases in the proportion of incomplete observations causes large reductions in the information about population parameters that is available

conditional on other variables in the model. There will be no problem with the assumption if the r observations have been deliberately chosen at random from the n. This is quite common in real world data sets when some variables are more expensive to measure than are others. Deliberate "double sampling" for sample surveys is described by Cochran (1963) with the objective of either maximising estimation precision for given financial resources or minimising the cost of attaining a specified precision. A large scale example of such a procedure is the data collection undertaken by the U.S. Bureau of the Census when collecting Census data.[2] Here each household receives either a short-form or a long-form. The long-form questionnaire includes the same 6 population questions (related to age, gender and marital status) and 1 housing question that are on the Census short-form, *plus* 26 additional population questions (including education, health, employment status and income) and 20 additional housing questions. On average about 1 in every 6 household received the long form and gives rise to exactly the data structure analysed in this paper.

Even in controlled randomised experiments the same motivation for limiting expensive variable measurement has repeatedly led to double sampling schemes (for example: Conniffe and Moran, 1972; Engel and Walstra, 1991; Caseur, 2005). The transparency property of our approach is particularly useful at the design stage of such observational studies. With an explicit formula for the variance of interest, the number of observations needed to attain a desired precision can be determined as can the optimal (in a cost minimisation sense) allocation between complete and incomplete observations.

The data structure of r complete and $(n–r)$ incomplete observations also arises frequently in econometrics through mechanisms other than deliberate random sampling. For example in many fields, such as labour economics, there is a growing tendency to draw data from multiple sources. This gives rise to a number of possibilities. It may be the case that the sample size differs between the two sources. Dolton and O'Neill (1996) presented an evaluation of a government training programme in the UK where data on personal

---

from the data. In their application almost all of the bounds are very large and span zero. In addition the computation of the bounds may be very time consuming.

characteristics such as sex, age, treatment status and some outcome data were obtained at the initial interview and design stage for the full sample of 8925. However other data, such as more detailed personal characteristics, previous employment history, search behaviour and data on non-labour income were obtained from a survey conducted 6 months later. This latter survey was completed by only 5200 of the original sample.

Even when the total sample sizes are the same in both data sources it is often the case that information obtained from one data source tends to be less prone to non response than that obtained from the second source. Possible examples include the use of linked employer-employee data sets (for a recent review see for Hamermesh (1999)) or the combination of administrative and survey data. In the former some firm related data such as tenure, wages and firm size may be completely measured for all respondents using firm-payroll data, whereas individual level data such as education and health are only available from the individual surveys and thus more likely to suffer from missing data issues. In the second example administrative data is often used to provide accurate measures of outcome variables such as earnings or unemployment histories, along with some limited personal data (often age and gender), while survey data are used to identify more detailed demographic characteristics such as education, marital status and family size. Examples include recent evaluations of the long-run effect of training programmes (Couch (1992) and Dolton and O'Neill (2002)). As with the linked employee-employer data non-response is more likely to occur with the survey, rather than administrative data, so that variables derived from this source may only be available for a subset of the entire sample. Researchers in this situation can either use the full sample restricted to the subset of variables obtained from the administrative data (as in Dolton and O'Neill (2002)) or use the full range of explanatory variables for the complete cases only. Neither approach is ideal.

In macroeconomics econometricians working with published official time series statistics can find that while all variables are available annually, some are also available quarterly. In some cases the recording of some variables may also have commenced well

[2] See for example the description of the U.S. Census 2000 at www.census.gov.

before that of others. Both situations could give rise to the type of data structure we analyse in this paper.

Furthermore, with our approach one can test the validity of the MAR assumption in cases where there is not deliberate double sampling. If it is true, coefficient estimates based on the complete data are consistent, but inefficient, while the estimates based on all data are consistent and efficient. If the two sets of estimates look very different the assumption is probably untrue. If the two sets of estimates are similar, with reduced standard errors for the estimates based on all data, the assumption is probably true. More formally, a Hausman (1978) type test can be performed based on the explicit variance formulae we derive. It is worth noting that should the test reject the assumption, the conclusion is not necessarily that inference should be based on the complete observation estimates. The implications for inference will depend greatly on which population is considered of real interest – that for which w is observable or the wider one. In the latter case, which is probably the norm in economics, the complete data is unrepresentative of the relevant population, so that the complete data estimates may be useless.[3]

In his 2006 presidential address to the American Finance Association, Campbell (2006) outlines the issues that arise when studying portfolio allocation decisions, noting in particular the data requirements for such analyses. In our application we use our estimator to examine the portfolio allocation decisions of Italian households using the Bank of Italy's Survey of Household Income and Wealth (SHIW). The SHIW data have been used to study a range of economic issues including wage risk and intertemporal labour supply (Pistaferri (2003)), schooling returns (Brunello and Miniaci (1999) and intertemporal choice and consumption mobility (Jappelli and Pistaferri (2006)). A major advantage of these data for the study of portfolio allocation is that they contain a question permitting estimation of a quantitative measure of risk-aversion. However, the question was only asked of a randomly chosen half of the total sample. This example is one whereby the majority of missing data is

ignorable by design and where complete case analysis involves dispensing with over half of the original sample. Using our estimator on the full data set produces standard errors that are approximately half those obtained under the complete case restriction. As a result a number of coefficients that were imprecisely estimated previously become significant. Such dramatic changes are a clear illustration of the potential gains which may be achieved by using all the data in an efficient manner.

The structure of the paper is as follows. Section 2 specifies the model and data structure we consider. Section 3 presents the efficient estimator for this model. Section 4 obtains explicit formulae for the asymptotic variance of our estimator, while Section 5 compares these results to the case where the dependent variable is continuous rather than a binary indicator. Section 6 presents some Monte Carlo simulations to assess the performance of our estimator. Section 7 presents the empirical application using the SHIW data and section 8 concludes. All proofs are provided in the Appendices.

2. Model Specification

We consider the following regression model :

$$Y_i = x_i' B_x + w_i' B_w + \varepsilon_i \tag{1}$$

where x and w are (k x 1) and (l x 1) vectors of regressors and $\varepsilon_i \sim N(0,1)$.[4] In addition:

$$w_i' = x_i' C + u_i' \tag{2}$$

where C is a (k x l) matrix of parameters, $u_i' \sim N(0,\Sigma)$ and (ε_i, u_i') are multivariate normally distributed (conditional on x).[5]

---

[3] In situations where the data are not MAR we say that the missing data mechanism are nonignorable. In this case the missing data mechanism must be modelled along with the substantive model. Examples include the sample-selection models considered by Heckman (1976, 1979).

We observe x, *w* and Z_i, where

$$Z_i = 1 \quad \text{if } Y_i > 0$$
$$Z_i = 0 \quad \text{if } Y_i \le 0 \tag{3}$$

The parameter vector to be estimated, θ, consists of the *k* components of B_x, the *l* components of B_w, the *l*k* elements of the matrix *C* and the $\left(\frac{l}{2}(l+1)\right)$ distinct elements of Σ.

We consider situations where data is available on $\{x_{i,}, w_i, Z_i\}$ for i=1….*r*. This represents the complete observation sample. In addition there are a further (*n*-*r*) observations on which $\{x_{i,}Z_i\}$ alone are measured. Complete case analysis estimates θ using only the observations i=1….*r*. In the next section we develop an efficient estimator for our data structure that makes use of the additional (*n*-*r*) observations.

3. Efficient Estimator

To derive our efficient estimator we use the fact that whenever $\tilde{\tilde{\theta}}$ is a $\sqrt{n}$ consistent for estimator for θ then the 'one-step' estimator

$$\hat{\theta} = \tilde{\tilde{\theta}} + J(\tilde{\tilde{\theta}})\frac{\partial}{\partial\theta}L_n(\tilde{\tilde{\theta}}) \tag{4}$$

where

$$J(\theta) = \left\{-\text{plim}\left[\frac{\partial^2}{\partial\theta\partial\theta'}L(\theta)\right]\right\}^{-1}$$

is asymptotically efficient (for example, Cox and Hinkley, 1974, p.308).

---

[4] The choice of a unit variance matches the conventional assumption of standard probit analysis and implies that the variance of Y conditional on x only is given by $\sigma_{yy} = 1 + B_w'\Sigma B_w$.

[5] Semiparametric approaches, such as Robins, Rotnitzky and Zhao (1994) and Robins. Hsieh and Newey (1995) relax the parametric assumptions concerning the covariate distribution. Although they do not consider the probit model explicitly, they show that their class of estimators contains an estimator whose asymptotic variance attains the semi-parametric variance bound for the models considered. Unfortunately this estimator may not be available for data analysis without further assumptions that are not required by our approach and even then their estimator may be difficult to implement.

Let $\tilde{\theta}' = (\tilde{B}_x', \tilde{B}_w', vec'\tilde{C}, vech'\tilde{\Sigma})$ denote the maximum likelihood estimator obtainable from the *r* complete observations. $\widetilde{B}_x$ and $\widetilde{B}_w$ are the coefficients from a standard probit analysis with *x* and *w* as explanatory variables, $\widetilde{C}$ is $(\widetilde{c}_1, \widetilde{c}_2, .\, \widetilde{c}_l)$, where $\widetilde{c}_j$ is the OLS coefficient vector for regression of the *j*th *w* on the *x* variables and $\widetilde{\Sigma}$ is the estimator of Σ based on the OLS residuals. As $\widetilde{\theta}$ is the ML estimator it is $\sqrt{r}$ consistent and therefore $\sqrt{n}$ consistent if we assume *n* proportional to *r*. Using (4) it follows that:

$$\hat{\theta} = \tilde{\theta} + J(\tilde{\theta})\frac{\partial}{\partial \theta} L_n(\tilde{\theta}) \tag{5}$$

is asymptotically efficient for θ.

The derivation of $\hat{\theta}$ requires the calculation of $J(\tilde{\theta})$ and $\frac{\partial}{\partial \theta} L_n(\tilde{\theta})$. For our data structure the log-likelihood function may be written

$$L_n = L_{r,z|w} + L_{r,w} + L_{n-r,z} \tag{6}$$

where the subscript *r* indicates complete observations and $(n-r)$ indicates incomplete observations. In Appendix A we use this to derive the required components of the efficient estimator (5). We show that efficient estimators for B_x and B_w are given by[6] :

$$\hat{B}_x = \widetilde{B}_x - \left[\widetilde{V}_x\left(\frac{\partial A}{\partial B_x}\right)_{\tilde{\theta}} + \widetilde{C}_{xw}\left(\frac{\partial A}{\partial B_w}\right)_{\tilde{\theta}}\right]\left(\overline{V}_{\tilde{A}} + \widetilde{V}_{\tilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right) \tag{7}$$

and

$$\hat{B}_w = \widetilde{B}_w - \left[\widetilde{C}_{xw}'\left(\frac{\partial A}{\partial B_x}\right)_{\tilde{\theta}} + \widetilde{V}_w\left(\frac{\partial A}{\partial B_w}\right)_{\tilde{\theta}}\right]\left(\overline{V}_{\tilde{A}} + \widetilde{V}_{\tilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right) \tag{8}$$

[6] Since this paper is primarily concerned with estimation of the coefficients of the probit regression of *Z* on *x* and *w* we focus on efficient estimators for B_x and B_w. However, the overall estimator $\hat{\theta}$, as given by (5) also provides efficient estimators of $\hat{C}$ and $\hat{\Sigma}$ of C and Σ. These are discussed briefly in the Appendix C.

where $A = \frac{(B_x + CB_w)}{\sqrt{1 + B_w^{'} \Sigma B_w}} = \frac{1}{\sqrt{\sigma_{yy}}}(B_x + CB_w)$, $\widetilde{A}$ results from A by replacing θ by $\widetilde{\theta}$.

and $\overline{A}$ is the MLE of A from (n-r) incomplete observations obtained from a simple probit of Z_i on X_i. $\tilde{V}_{\tilde{A}}$ and $\tilde{V}_{\overline{A}}$ denote their estimated variance matrices. Likewise $\tilde{V}_x$ and $\tilde{V}_w$ denote the variance-covariance matrices of $\widetilde{B}_x$ and $\widetilde{B}_w$, respectively, evaluated at the MLE estimates and $\tilde{C}_{xw}$ their estimated covariance matrix.

Consistency of $\hat{B}_x$ and $\hat{B}_w$ requires that $\widetilde{A} - \overline{A}$ be a consistent estimator of zero. A necessary condition for this is that the missing data for *w* are MAR. As noted earlier this is a common assumption in much of the work on missing data and in the next section we will show how to test validity of the assumption within our framework. Should the test imply that the MAR assumption is false, the implications for inference will depend greatly on which population is considered of real interest – that for which *w* is observable or the wider population including those who cannot or will not provide *w*. In the former case, B_x is estimable from the complete data, but no use can be made of the extra data. In the latter case, the complete data may be unrepresentative of the wider population, so that the complete data estimate may be of little use.

4. Asymptotic Variance:

The asymptotic variances of $\hat{B}_x$ and $\hat{B}_w$ are derived in the Appendix B. To do this we note that since $\hat{B}_x$ and $\widetilde{B}_x - \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\overline{A}} + V_{\tilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)$ differ only in terms of $O_p(n^{-1})$ they have the same asymptotic variance. It can then be shown that:

$$(9) \qquad Var(\hat{B}_x) = V_x - \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\overline{A}} + V_{\tilde{A}}\right)^{-1}\left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]^{'}$$

and the estimated variance is obtained by replacing the Vs by $\tilde{V}s$, C_{xw} by $\tilde{C}_{xw}$ and the derivates by their values evaluated at $\tilde{\theta}$.

Similarly, the variance of $\hat{B}_w$ may be shown to be

$$(10) \qquad Var(\hat{B}_w) = V_w - \left[C'_{xw}\left(\frac{\partial A}{\partial B_x}\right) + V_w\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\left[C'_{xw}\left(\frac{\partial A}{\partial B_x}\right) + V_w\left(\frac{\partial A}{\partial B_w}\right)\right]'$$

and the covariance of $\hat{B}_x$ and $\hat{B}_w$ to be

$$(11) \qquad Cov(\hat{B}_x, \hat{B}_w) = C_{xw} - \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\left[C'_{xw}\left(\frac{\partial A}{\partial B_x}\right) + V_w\left(\frac{\partial A}{\partial B_w}\right)\right]'.$$

5. The Case of Observed *Y*.

Before examining our estimator in more detail we first briefly discuss its relationship to some earlier estimators developed in the literature. The estimators given in (7) and (8) followed from the structure of the likelihood given by equation (6). Since this structure is not unique to probit regression similar estimators exist for other models. An obvious case is that of observed *Y* with the same assumptions about the relationships between *Y*, the *w* variables and the *x* variables. In Appendix D we show that for this model

$$(12) \qquad \hat{B}_x = \tilde{B}_x - \frac{\tilde{\sigma}_{yy.w}}{\tilde{\sigma}_{yy}}\tilde{V}_{\tilde{A}}\left(\bar{V}_{\bar{A}} + \tilde{V}_{\tilde{A}}\right)^{-1}\left(\tilde{A} - \bar{A}\right)$$

and

$$(13) \qquad \hat{B}_w = \tilde{B}_w .$$

where $\widetilde{B}_x$ and $\widetilde{B}_w$ are now the usual OLS estimators, $\widetilde{A}$ and $\overline{A}$ are OLS estimators of coefficients of *Y* on just the *x* variables for the *r* and *(n– r)* observations respectively, $\sigma_{yy.w}$ is simply estimated from the error mean square of the regression of *Y* on the *x* and *w* variables for the *r* complete observations, and σ_{yy} is estimated from the error mean square of regression of *Y* on *x* alone for the full *n* observations.

The failure to improve on $\widetilde{B}_w$ is intuitively plausible since the *w* variables are only measured on the *r* complete observations, while the *x* variables are measured on all *n*. While in the probit case the efficient estimator given in (8) is not identical to $\widetilde{B}_w$, we should not expect its variance to be much different from that of $\widetilde{B}_w$.

Introducing $\hat{A}$ as the efficient estimator of *A* over all *n* observations, obtained by weighting $\widetilde{A}$ and $\overline{A}$ inversely by their variances, it is easily shown that

$$\hat{B}_x = \widetilde{B}_x - \frac{\widetilde{\sigma}_{yy.w}}{\widetilde{\sigma}_{yy}}\left(\widetilde{A} - \overline{A}\right), \tag{14}$$

with asymptotic variance

$$V_x + \frac{\sigma^2_{yy.w}}{\sigma_{yy}}\left[(X_r' X_r)^{-1} - (X_n' X_n)^{-1}\right], \tag{15}$$

where X_r is the $(r \times k)$ matrix of *x* values for the complete observations and X_n is the $(n \times k)$ matrix of all *x* values. This is the estimator obtained in Conniffe (1983), which was shown to also have desirable finite sample properties as well as being asymptotically efficient. In particular, $\hat{B}_x$ is unbiased and an explicit exact finite sample variance is available. The earlier results were not derived from the likelihood function, as in this paper, but from the device, going back at least to Rao (1967), of modifying a consistent estimator $\widetilde{\theta}$ through

$$\hat{\theta} = \widetilde{\theta} + \Omega\, S \tag{16}$$

where S is a statistic correlated with $\tilde{\theta}$ and with asymptotic expectation zero. For any particular S the $\hat{\theta}$ with minimum variance is obtained by taking Ω equal to minus the covariance of $\widetilde{\theta}$ and S by the inverse of the variance of S. Choosing $S = \widetilde{A} - \hat{A}$ leads to the estimator $\hat{B}_x$. The device has been reapplied in recent papers, for example, by Chen and Chen (2002) in a partially semi-parametric context.

6. Simulations.

Before studying the determinants of portfolio allocation using the Bank of Italy's SHIW, we assess the performance of our estimator using Monte Carlo simulations. The model used for the simulations is

$$Y_i = x_i' B_x + w_i' B_w + \varepsilon_i \tag{17}$$

where *x* and *w* are both scalar random variables and $\varepsilon_i \sim N(0,1)$. For the simulation we assume that $x_i' \sim N(0,1)$. In addition:

$$w_i' = x_i' C + u_i' \tag{18}$$

where $u_i' \sim N(0,\sigma)$.

The true parameter vector θ', is therefore a (1x4) vector consisting of (B_x, B_w, C, σ). For the simulation we set θ'=(1,1,1,1).

We observe *x*, *w* and *Z*, where

$$\begin{aligned} Z_i &= 1 \quad \text{if } Y_i > 0 \\ Z_i &= 0 \quad \text{if } Y_i \le 0 \end{aligned} \tag{19}$$

We consider situations where data is available on $\{x_{i,}, w_i, Z_i\}$ for i=1….*r*. This represents the complete observation sample. In addition there are a further (*n*-*r*) observations on which $\{x_{i,}, Z_i\}$ alone are measured. The simulations ensure that the data are missing completely at random. We carry out the simulations for three different choices of *n* (500, 1000 and 6000) and also consider vary the proportion of missing data across the samples. In particular for each *n* we consider values of $\frac{(n-r)}{n}$ equal to .7, .5, .25 and .1.

The results of the simulations, based on 1000 replications, are given in Table 1.[7] The first four columns correspond to the point estimates and variances from the complete case analysis. The second four columns present the corresponding results using our efficient estimator. The results for the point estimates are as expected. There appears to be a small bias in the parameters that goes to zero as $r \to \infty$.[8] As expected there are no significant differences between the estimates across the two estimators and the true parameter vector is not rejected in any of the nine simulations.

However, when we turn to the estimated variances we see significant improvements in precision when the efficient estimator used. The results are consistent across sample sizes. In keeping with the findings from the linear regression model there is very little difference in the estimated variance of B_w. The failure to improve on $\widetilde{B}_w$ is intuitively plausible since the *w* variables are only measured on the *r* complete observations. However, a comparison of the estimated variances of $\hat{B}_x$ and $\tilde{B}_x$ show significant improvements in precision. As expected the biggest reductions in variance arise when the proportion of missing data is highest. In the worst case scenario considered, when 70% of the data are missing, we see an approximate sixty percent reduction in the variance. Even in cases with more moderate degrees of missing data the reductions in the estimated variance are non-trivial. The reduction in variance is of

[7] The estimates for our new estimator are easily obtained form a new user-written Stata package provided by the authors. This program, called *probitmiss*, along with a help file is available for download at http://economics.nuim.ie/staff/oneill/probitmissprograms.shtml.

[8] This is to be expected as the standard complete case Probit estimator is biased, as are maximum likelihood estimators in general.

the order of ten to twenty percent when we consider missing data in the range of ten to twenty five percent of the initial sample.[9]

Table 1 also allows us to compare the performance of our estimator to a popular multiple imputation technique for handling missing data. In columns 5-8, underneath the estimates from our efficient estimator, we present Monte Carlo results using the multiple imputation package provided in Stata (see Royston (2004)). This package imputes values for missing data by drawing imputations at random from the posterior distribution of the missing values of w, conditional on the observed values and the variables in $\{Z,x\}$. The results reported in Table 1 suggest that estimates and standard errors produced by the multiple implementation package are consistent with our efficient estimator when the proportion of missing data is small. However, the performance of the multiple imputation procedure becomes less satisfactory as the proportion of missing data rises. While our estimator remains effectively unbiased as the proportion of missing data increases, the estimator based on multiple imputation does not, with the bias increasing as the proportion of missing data increases.[10]

7. Empirical Application to Portfolio Allocation.

Campbell (2006) presents an overview of recent theoretical and empirical developments in the area of household financial decision making, noting that empirical studies in this field often encounter difficulties obtaining the high-quality data necessary. In this section we apply the results developed in the previous sections to look at the portfolio allocation decisions of Italian households using the Bank of Italy's Survey of Household Income and Wealth

[9] Other simulations, not presented, suggest that the improvements in efficiency increase as the correlation between x and w falls and as B_w decreases. These findings are intuitive and consistent with the results for the linear regression model (Conniffe (1983)).

[10] Paul et al (2008) report biases of similar magnitude to us when applying multiple imputation techniques to a logistic model. It is interesting to note that in our simulation the bias in the multiple imputation is only evident with the binary dependent variable. When Y_i is assumed to be fully observed, resulting in the standard linear regression model, the multiple imputation approach appears to be unbiased even when the degree of missing data is large.

(SHIW). The SHIW has been used recently to study issues such as the schooling returns in Italy (Brunello and Miniaci 1999), earnings and employment risk (Guiso et al 2002), wage risk and intertemporal labour supply (Pistaferri 2003) and intertemporal choice and consumption mobility (Jappelli and Pistaferri 2006). In the next section we discuss the strengths of the SHIW for studying portfolio allocation. We outline the problems of missing data that arise in this application and use our proposed estimator to examine the decision to hold risky assets. The application is used to illustrate the efficiency gains arising from our estimator relative to the traditional complete case analysis.

7.1 Bank of Italy's Survey of Household Income and Wealth

Since 1962, the Bank of Italy has conducted surveys on household budgets, which allows researchers to examine economic behavior at the micro level. The primary aim of the survey is to collect detailed information on income and savings of households. Campbell (2006) argues that an ideal data set for studying household financial decision making should meet five criteria; it should cover a representative sample of the entire population, should contain measures of total wealth, should identify individual assets so that one could measure household diversification, should be reported with a high-level of accuracy and should follow households over time. The SHIW performs well on each of these measures, being a repeated nationally representative sample of approximately 8000 Italian households, with finely disaggregated data on assets and wealth that are measured with reasonable accuracy.[11]

In addition to traditional measurement problems, previous studies of portfolio allocation have been limited by the extent to which they can measure risk-aversion. An important feature of the SHIW in this respect is that the later surveys contained questions that attempt to directly measure individual levels of risk-aversion. Both the 1995 and 2000

---

[11] The main purpose of this section example is to illustrate the efficiency gains arising from our new estimator. Other potential biases such as that from measurement error are not addressed directly. Biancotti et al (2008) provide a detailed analysis of measurement error issues in the SHIW. While there is variation in the reliability index across disaggregated assets overall the SHIW performed well.

surveys asked individuals to value a hypothetical lottery so as to measure their degree of risk aversion. The wording of the question varied slightly between surveys, so for clarity we focus only on the 2000 survey. In that year the lottery question was as follows:

"You are offered the opportunity of buying shares which, tomorrow, with equal probability, will be worth either **10 million** or **nothing**. How much would you be prepared to pay **(maximum amount)** to buy these shares?"

Thus individuals who pay *P* lire for this lottery have a 50% chance of winning (10m) and a 50% chance of winning zero. The expected value of this lottery net of the purchase price is .5*10m-*P*. Clearly individuals who are risk neutral will pay anything up to 5m to play this lottery, since the expected value of the winnings will still be positive. A risk-averse decision taker will pay less than 5m and a risk-lover would be willing to pay more than 5m lire. Using a Taylor series approximation of a utility function we obtain the following approximate expression for the Arrow-Pratt measure of absolute risk aversion[12]:

$$R_i(y) = \frac{(5 - P_i)}{[\frac{P_i^2}{2} + .5 * \frac{10^2}{2} - 5 * P_i]} \tag{17}$$

For individuals who are risk neutral P_i=5, so that $R_i(y)$=0.

However, there are two data problems associated with the lottery question in the SHIW. Firstly in 2000 it was only asked of a random sample of one half of the survey. In terms of the structure of our missing data problem, this is an ideal scenario in that by construction the data are missing at random. However on top of this we also have a problem of non-response by those scheduled to answer the question. In total the inclusion of the risk-aversion question reduces the sample size from 6779 to 1029. A traditional approach to

---

Reliability indices for the disaggregated income and wealth measured were typically over 70%, while the index for aggregate measures of net disposable income and net wealth was over 80%.

[12] See also Hartog et al (2002).

estimating this model would be to focus on the complete data. However in our application this involves throwing away over 5000 observations. The estimator proposed in our paper provides a way of incorporating these additional observations to improve the precision of the traditional estimator.

Table 2 presents descriptive statistics for the main variables used in our analysis. The dependent variable in our analysis is a binary variable indicating whether or not the household held risky assets as part of their savings portfolio at the end of 2000. The sample is restricted to those who reported positive savings as of the end of 2000. This leaves us with a base sample size of 6779. As noted earlier restricting ourselves to households with a valid measure of risk-aversion reduces our sample to 1029. Column one reports summary statistics for the base sample, while column 2 reported the summary figures for the subsample for which we can measure risk-aversion. Looking at the base sample we see that 23.5% of the sample report holding risky assets as part of their savings portfolio.[13] The average age of head of household was 54, while the proportion with college education was 10.3%. 31.5% of the household heads were women and 71% were married. The results for the subsample are given in column 2. The summary measures are broadly consistent with the full-sample, though they are some differences on the region variable. We will return to this issue when testing the validity of our missing at random assumption.

7.2 Estimation Results

Table 3 reports the results from our estimated model. The results for the complete case analysis are presented in the first two columns while the estimates based on the efficient estimator are given in the final two columns. Looking first at the results for the complete case analysis we see that as expected the greater the degree of risk-aversion the less likely it is that a household will hold risky assets in their portfolio. In addition older individuals and those

[13] Risky assets are defined as bonds, shares of Italian mutual funds or equity. Non risky assets include deposit accounts and government securities.

with a college education are also more likely to hold risky assets.[14] Those located in the south or the islands are less likely to hold risky assets.[15] Of the remaining coefficients neither the gender, marital status or the North-West or Centre region variables are precisely estimated for the complete sample case.

Columns three and four report the results from the efficient estimator developed in this paper. The fact that the point estimates from the efficient estimator are comparable to those from the complete case analysis supports our assumption of missing at random. This assumption can be tested through a Hausman (1978) type test. Under the assumption of MAR $\hat{B}_x$ is the efficient estimator, with variance given by (9) and $\widetilde{B}_x$ is a consistent estimator, with variance V_x. Since the asymptotic variance of the difference between an efficient estimator and another consistent one is the difference of the variances, then:

$$(\hat{B}_X - \widetilde{B}_X)' \left\{ \left[\widetilde{V}_X \left(\frac{\partial A}{\partial B_X} \right)_{\widetilde{\theta}} + \widetilde{C}_{XW} \left(\frac{\partial A}{\partial B_W} \right)_{\widetilde{\theta}} \right] \left(\overline{V}_{\overline{A}} + \widetilde{V}_{\widetilde{A}} \right)^{-1} \left[\widetilde{V}_X \left(\frac{\partial A}{\partial B_X} \right)_{\widetilde{\theta}} + \widetilde{C}_{XW} \left(\frac{\partial A}{\partial B_W} \right)_{\widetilde{\theta}} \right]' \right\}^{-1} (\hat{B}_X - \widetilde{B}_X)$$

is asymptotically χ^2 with k degrees of freedom. Applying this test to our application leads to a χ^2 statistic 11.39, with an associated p-value of .25, which supports the assumptions underlying our estimator for this application.

Having tested the underlying assumptions of our estimator we can now look at the efficiency gains achieved from our approach. A comparison of the standard errors across the two estimators shows substantial efficiency gains from the new estimator. For almost all the parameters the standard errors from the efficient estimator are half those of the complete case analysis. The exception is the coefficient on risk-aversion for which the standard error is virtually the same. This is to be expected since the extra data used in the efficient estimator contains no independent information on risk-aversion. However, for the other variables the standard errors have been reduced significantly. The result is that explanatory variables such

[14] These results are consistent with previous studies of portfolio allocation (e.g Guiso et al (1996) and Rosen and Wu (2004)) though neither of those studies directly controlled for individual risk-aversion.
[15] The omitted region refers to those living in the North-East.

as marital status, the north-west dummy and the central regional dummy, which were insignificant in the complete case analysis, are now precisely estimated with coefficients that are similar to those from the complete case analysis.

8. Conclusion

In this paper we develop an asymptotically efficient estimator for handling missing data on explanatory variables in a probit choice model, that is easily implemented using standard software packages such as Stata. We provide closed form expressions for both the estimator and its asymptotic variance and relate these to previous results obtained for the case where the dependent variable is continuous rather than binary. We also carry out simulations which illustrate that our estimator outperforms popular alternative approaches.

In our application we use our estimator to study the portfolio allocation decision of Italian households using the Bank of Italy's SHIW data. In this situation complete case analysis results in over half of the data being discarded. A Hausman test is used to check the validity of the ignorable data assumption underlying our estimator, while use of the efficient estimator leads to standard errors that are, in most cases, half the size of those obtained using only the complete cases. As a result a number of coefficients that were imprecisely estimated previously are now significant.

The substantial improvement in precision arising from our estimator, the transparency provided by the closed form expressions for the estimator and its variance and the ease with which the estimator can be implemented provides an attractive new option for binary choice analysis with missing data.

Appendix A: **Efficient estimators of B_x and B_w**

As noted in the main text our data structure implies that the log-likelihood function over the entire sample L_n may be written as

(A1) $$L_n = L_{r,z|w} + L_{r,w} + L_{n-r,z}$$

where the subscript r indicates complete observations and $(n-r)$ indicates incomplete observations.

Under our normality assumptions the first component of the likelihood based on the complete observations is

$$L_{r,z|w} = \sum_{1}^{r} \{z_i \log \Phi(M_i) + (1-z_i)\log[1-\Phi(M_i)]\},$$

with

$$M_i = x_i' B_x + w_i' B_w .$$

The second is

$$L_{r,w} = -\frac{\mathrm{rl}}{2}\log 2\pi - \frac{r}{2}\log|\Sigma| - \frac{1}{2}\sum_{1}^{r}(w_i - C'x_i)'\Sigma^{-1}(w_i - C'x_i),$$

which is the likelihood function for a seemingly unrelated regressions model with the same explanatory variables in each equation. The third is

$$L_{n-r} = \sum_{r+1}^{n} \{z_i \log \Phi(M_i^*) + (1-z_i)\log[1-\Phi(M_i^*)]\},$$

with

$$M_i^* = \frac{x_i'(B_x + CB_w)}{\sqrt{1 + B_w'\Sigma B_w}} = x_i' A .$$

The k element vector A is the unconditional (or conditionally on x alone) mean of the underlying unobserved Y divided by its unconditional standard error. The vector of all parameters, θ, is the transpose of $\theta' = [B_x', B_w', vec'C, vech'\Sigma]$, where *vech* denotes the half-vectorization operator that transforms a symmetric matrix into a vector, omitting the

duplicated elements above the leading diagonal (see for example Seber (2008)). In total, there are $q = k + l + kl + l(l+1)/2$ parameters.

Derivation of the efficient estimator requires the calculation of $J(\theta)$ and $\frac{\partial}{\partial\theta}L_n(\theta)$ evaluated at $\tilde{\theta}' = (\tilde{B}_x', \tilde{B}_w', vec'\tilde{C}, vech'\tilde{\Sigma})$, the maximum likelihood estimator of θ using only the *r* complete observations. Since $L_r = L_{r,z|w} + L_{r,w}$ the $\widetilde{B}_x$ and $\widetilde{B}_w$ are independent of $\widetilde{C}$ and $\widetilde{\Sigma}$.

Remembering that A is a function of θ

$$\frac{\partial L}{\partial\theta} = \frac{\partial L_r}{\partial\theta} + \frac{\partial A}{\partial\theta}\frac{\partial L_{n-r}}{\partial A},$$

and so

$$\left(\frac{\partial L}{\partial\theta}\right)_{\tilde{\theta}} = \left(\frac{\partial L_r}{\partial\theta}\right)_{\tilde{\theta}} + \left(\frac{\partial A}{\partial\theta}\right)_{\tilde{\theta}}\left(\frac{\partial L_{n-r}}{\partial A}\right)_{\tilde{A}},$$

where $\widetilde{A}$ results from A by replacing θ by $\widetilde{\theta}$. It is worth emphasising that $\widetilde{A}$ is not the same as the estimates from a standard probit regression on the *x* variables for the *r* complete observations, denoted bys A^*. Chesher (1984) compared $\widetilde{A}$ with A^* in the context of estimating a probit equation on *x* variables jointly with a linear regression of *w* on the same variables in a seemingly unrelated regression system, assuming bivariate normality of the unobserved *Y* and *w*. He concluded A^* could be very inefficient and was supported by Ronning and Kukuk (1996).

Denoting the MLE of A from $L_{n-r}(A)$ by $\overline{A}$

$$\left(\frac{\partial L_{n-r}}{\partial A}\right)_{\tilde{A}} = \left(\frac{\partial L_{n-r}}{\partial A}\right)_{\overline{A}} + \left(\frac{\partial^2 L_{n-r}}{\partial A\partial A'}\right)_{\overline{A}}(\widetilde{A} - \overline{A}) + O_p(1).$$

The derivative of L_{n-r} is zero at $\overline{A}$ and

$$-\left(\frac{\partial^2 L_{n-r}}{\partial A\partial A'}\right)^{-1}_{\overline{A}} = \overline{V}_{\overline{A}},$$

which estimates $V_{\bar{A}}$, the variance of $\bar{A}$, and satisfies $\bar{V}_{\bar{A}} = V_{\bar{A}} + O_p(n^{-\frac{3}{2}})$. So

(A2) $$\left(\frac{\partial L}{\partial \theta}\right)_{\tilde{\theta}} = -\left(\frac{\partial A}{\partial \theta}\right)_{\tilde{\theta}} \bar{V}_{\bar{A}}^{-1}(\tilde{A} - \bar{A}) + O_p(1)$$

which is $O_p(\sqrt{n})$.

Turning to the second derivative

$$\frac{\partial^2 L}{\partial \theta \partial \theta'} = \frac{\partial^2 L_r}{\partial \theta \partial \theta'} + \left(\frac{\partial L_{n-r}}{\partial A} \otimes I_q\right)\left(\frac{\partial}{\partial \theta} vec \frac{\partial A}{\partial \theta}\right) + \left(\frac{\partial A}{\partial \theta}\right)\frac{\partial}{\partial \theta}\left(\frac{\partial L_{n-r}}{\partial A}\right)$$

and

$$\frac{\partial}{\partial \theta}\left(\frac{\partial L_{n-r}}{\partial A}\right) = \frac{\partial^2 L_{n-r}}{\partial A \partial A'}\left(\frac{\partial A}{\partial \theta}\right)'.$$

So

$$\left(\frac{\partial^2 L}{\partial \theta \partial \theta'}\right)_{\tilde{\theta}} = \left(\frac{\partial^2 L_r}{\partial \theta \partial \theta'}\right)_{\tilde{\theta}} + \left[\left(\frac{\partial L_{n-r}}{\partial A}\right)_{\tilde{A}} \otimes I_q\right]\left(\frac{\partial}{\partial \theta} vec \frac{\partial A}{\partial \theta}\right)_{\tilde{\theta}} + \left(\frac{\partial A}{\partial \theta}\right)_{\tilde{\theta}}\left(\frac{\partial^2 L_{n-r}}{\partial A \partial A'}\right)_{\tilde{A}}\left(\frac{\partial A}{\partial \theta}\right)'_{\tilde{\theta}}.$$

Now

$$\left(\frac{\partial^2 L_r}{\partial \theta \partial \theta'}\right)_{\tilde{\theta}} = -\tilde{V}_{\tilde{\theta}}^{-1} = -V_{\tilde{\theta}}^{-1} + O_p(\sqrt{n}),$$

where $V_{\tilde{\theta}}$ is the variance matrix of $\tilde{\theta}$, the MLE of θ from $L_r(\theta)$, estimated by $\tilde{V}_{\tilde{\theta}}$. Also

$$\left(\frac{\partial^2 L_{n-r}}{\partial A \partial A'}\right)_{\tilde{A}} = \left(\frac{\partial^2 L_{n-r}}{\partial A \partial A'}\right)_{\bar{A}} + O_p(\sqrt{n}) = -\bar{V}_{\bar{A}}^{-1} + O_p(\sqrt{n})$$

and

$$\left(\frac{\partial L_{n-r}}{\partial A}\right)_{\tilde{A}} = -\bar{V}_{\bar{A}}^{-1}(\tilde{A} - \bar{A}) + O_p(1)$$

is $O_p\left(\sqrt{n}\right)$ while

$$\left(\frac{\partial}{\partial \theta} vec \frac{\partial A}{\partial \theta}\right)_{\tilde{\theta}}$$

is $O_p(1)$. So

$$\left(\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right)_{\tilde\theta} = -\left[\tilde V_{\tilde\theta}^{-1} + \left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta} \bar V_{\bar A}^{-1}\left(\frac{\partial A}{\partial\theta}\right)'_{\tilde\theta}\right] + O_p(\sqrt{n}).$$

Using the matrix inversion formula

$$(R+STU)^{-1} = R^{-1} - R^{-1}S\left(T^{-1}+UR^{-1}S\right)^{-1}UR^{-1}$$

gives

$$\left(\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right)^{-1}_{\tilde\theta} = -\left\{\tilde V_{\tilde\theta} - \tilde V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta}\left[\bar V_{\bar A} + \left(\frac{\partial A}{\partial\theta}\right)'_{\tilde\theta}\tilde V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta}\right]^{-1}\left(\frac{\partial A}{\partial\theta}\right)'_{\tilde\theta}\tilde V_{\theta}\right\} + O_p(n^{-\frac{3}{2}})$$

.

Since A is a function of θ, the asymptotic variance of $\tilde A$ is

$$V_{\tilde A} = \left(\frac{\partial A}{\partial\theta}\right)' V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right),$$

which is estimated by

$$\tilde V_{\tilde A} = \left(\frac{\partial A}{\partial\theta}\right)'_{\tilde\theta} \tilde V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta}$$

and so

(A3) $$\left(\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right)^{-1}_{\tilde\theta} = -\left\{\tilde V_{\tilde\theta} - \tilde V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta}\left[\bar V_{\bar A} + \tilde V_{\tilde A}\right]^{-1}\left(\frac{\partial A}{\partial\theta}\right)'_{\tilde\theta}\tilde V_{\tilde\theta}\right\} + O_p(n^{-\frac{3}{2}}).$$

Using (A2) and (A3)

$$\left(\frac{\partial^2 L}{\partial\theta\,\partial\theta'}\right)^{-1}_{\tilde\theta}\left(\frac{\partial L}{\partial\theta}\right)_{\tilde\theta} = \left\{\tilde V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta}\left[I - \left(\bar V_{\bar A} + \tilde V_{\tilde A}\right)^{-1}\right]\tilde V_{\tilde A}\right\}\bar V_{\bar A}^{-1}\left(\tilde A - \bar A\right) + O_p(n^{-1})$$

(A4) $$= \widetilde{V}_{\tilde{\theta}} \left(\frac{\partial A}{\partial \theta} \right)_{\tilde{\theta}} \left(\overline{V}_{\bar{A}} + \widetilde{V}_{\tilde{A}} \right)^{-1} \left(\widetilde{A} - \overline{A} \right) + O_p(n^{-1}) .$$

Using (A4) the efficient estimator, $\breve{\theta} = \widetilde{\theta} + J(\widetilde{\theta}) \frac{\partial}{\partial \theta} L_n(\widetilde{\theta})$, can be written as

$$\breve{\theta} = \widetilde{\theta} - \widetilde{V}_{\tilde{\theta}} \left(\frac{\partial A}{\partial \theta} \right)_{\tilde{\theta}} \left(\overline{V}_{\bar{A}} + \widetilde{V}_{\tilde{A}} \right)^{-1} \left(\widetilde{A} - \overline{A} \right) + O_p(n^{-1})$$

and since estimators differing only in a term of $O_p(n^{-1})$ have the same asymptotic variance

(A5) $$\hat{\theta} = \widetilde{\theta} - \widetilde{V}_{\tilde{\theta}} \left(\frac{\partial A}{\partial \theta} \right)_{\tilde{\theta}} \left(\overline{V}_{\bar{A}} + \widetilde{V}_{\tilde{A}} \right)^{-1} \left(\widetilde{A} - \overline{A} \right)$$

is an efficient estimator. Denoting the variance matrix of $\widetilde{B}_x$ by V_x, that of $\widetilde{B}_w$ by V_w and their covariance by C_{xw},

(A6) $$V_{\tilde{\theta}} = \begin{bmatrix} V_x, & C_{xw}, & 0, & 0 \\ C_{xw}', & V_w, & 0, & 0 \\ 0, & 0, & \Sigma \otimes (X'X)^{-1}, & 0 \\ 0, & 0, & 0, & H \end{bmatrix},$$

where $\Sigma \otimes (X'X)^{-1}$ is the variance matrix of $vec\,\widetilde{C}$, the $k*l$ vector of coefficients from OLS regressions of w variables on x and H is the variance matrix of the $l(l+1)/2$ element vector of OLS estimates of the lower triangular components of Σ. The elements of H are of the form $(\sigma_{ij}\sigma_{i^*j^*} + \sigma_{ij^*}\sigma_{ji^*})/r$ as is shown in standard textbooks (e.g. Kendall and Stuart, vol. 3, pg 254).

$\widetilde{V}_{\tilde{\theta}}$, the estimator of $V_{\tilde{\theta}}$, is obtained by replacing V_x, V_w and C_{xw} in (A6) by $\widetilde{V}_x, \widetilde{V}_w$ and $\widetilde{C}_{xw}$ respectively, where these are produced by the standard probit regression for

the *r* complete observations, and Σ and H by $\widetilde{\Sigma}$ and $\widetilde{H}$, where the σ_{ij} are replaced by their estimators based on OLS residuals. From the structure of (A6) it is clear that

$$\text{(A7)} \qquad \hat{B}_x = \widetilde{B}_x - \left[\widetilde{V}_x\left(\frac{\partial A}{\partial B_x}\right)_{\widetilde{\theta}} + \widetilde{C}_{xw}\left(\frac{\partial A}{\partial B_w}\right)_{\widetilde{\theta}}\right]\left(\overline{V}_{\widetilde{A}} + \widetilde{V}_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)$$

and

$$\text{(A8)} \qquad \hat{B}_w = \widetilde{B}_w - \left[\widetilde{C}'_{xw}\left(\frac{\partial A}{\partial B_x}\right)_{\widetilde{\theta}} + \widetilde{V}_w\left(\frac{\partial A}{\partial B_w}\right)_{\widetilde{\theta}}\right]\left(\overline{V}_{\widetilde{A}} + \widetilde{V}_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right).$$

These are the expressions that appear in equations (7) and (8) of the main text.

For completeness we note that since

$$A = \frac{(B_x + CB_w)}{\sqrt{1 + B_w'\Sigma B_w}} = \frac{1}{\sqrt{\sigma_{yy}}}(B_x + CB_w)$$

it is clear that

$$\frac{\partial A}{\partial B_x} = \frac{1}{\sqrt{\sigma_{yy}}} I_k$$

and

$$\frac{\partial A}{\partial B_w} = \frac{1}{\sqrt{\sigma_{yy}}} C' - \frac{1}{2\sigma_{yy}}\frac{\partial \sigma_{yy}}{\partial B_w} A' = \frac{1}{\sqrt{\sigma_{yy}}} C' - \frac{1}{\sigma_{yy}} \Sigma B_w A'.$$

So

$$\text{(A9)} \qquad \left(\frac{\partial A}{\partial B_x}\right)_{\widetilde{\theta}} = \frac{1}{\sqrt{\widetilde{\sigma}_{yy}}} I_k \text{ and } \left(\frac{\partial A}{\partial B_w}\right)_{\widetilde{\theta}} = \frac{1}{\sqrt{\widetilde{\sigma}_{yy}}} \widetilde{C}' - \frac{1}{\widetilde{\sigma}_{yy}} \widetilde{\Sigma}\widetilde{B}_w \widetilde{A}',$$

where $\widetilde{\sigma}_{yy} = 1 + \widetilde{B}_w'\widetilde{\Sigma}\widetilde{B}_w$. It may be worth noting that ΣB_w is the vector of 'covariances' of the unobserved *Y* and the *w* variables (conditionally on the *x* variables).

Appendix B: **Variances of $\hat{B}_x$ and $\hat{B}_w$**

To obtain the variance of $\hat{B}_x$ as given by (A7) note that

$$\widetilde{V}_x\left(\frac{\partial A}{\partial B_x}\right)_{\tilde{\theta}} + \widetilde{C}_{xw}\left(\frac{\partial A}{\partial B_w}\right)_{\tilde{\theta}} = V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right) + O_p(n^{-\frac{3}{2}})$$

and

$$\left(\overline{V}_{\overline{A}} + \widetilde{V}_{\widetilde{A}}\right)^{-1} = \left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1} + O_p(\sqrt{n}).$$

Bearing in mind that $\widetilde{A} - \overline{A}$ is $O_p(n^{-\frac{1}{2}})$ it follows that $\hat{B}_x$ and

$$\widetilde{B}_x - \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)$$

differ only in terms of $O_p(n^{-1})$ and so have the same asymptotic variance. Letting

$$\Lambda = \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]$$

and remembering that $\widetilde{B}_x$ and $\overline{A}$ are independent, the (asymptotic) variance of $\hat{B}_x$ is

$$E\left\{\left[\widetilde{B}_x - \Lambda\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)\right]\left[\widetilde{B}_x - \Lambda\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)\right]'\right\} - B_x B_w$$

$$= V_x - \operatorname{cov}(\widetilde{B}_x, \widetilde{A})\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\Lambda' - \Lambda\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\operatorname{cov}(\widetilde{A}, \widetilde{B}_x) + \Lambda\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\Lambda'$$

having used the fact that the variance of $\widetilde{A} - \overline{A}$ is $V_{\overline{A}} + V_{\widetilde{A}}$. Since

$$\widetilde{A} = A + \frac{\partial A}{\partial B_x}(\widetilde{B}_x - B_x) + \frac{\partial A}{\partial B_w}(\widetilde{B}_w - B_w) + (\text{terms independent of } \widetilde{B}_x) + O_p(n^{-1})$$

the covariance of $\widetilde{B}_x$ and $\widetilde{A}$ is Λ. Therefore the variance of $\hat{B}_x$ is

$$Var(\hat{B}_x) = V_x - \Lambda\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\Lambda'$$

(A10)
$$= V_x - \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\overline{A}} + V_{\widetilde{A}}\right)^{-1}\left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]'$$

This is equation (9) in the main text.

Similarly, the variance of $\hat{B}_w$ may be shown to be

$$V_w - \left[C'_{xw}\left(\frac{\partial A}{\partial B_x}\right) + V_w\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\left[C'_{xw}\left(\frac{\partial A}{\partial B_x}\right) + V_w\left(\frac{\partial A}{\partial B_w}\right)\right]' \tag{A11}$$

and the covariance of $\hat{B}_x$ and $\hat{B}_w$ to be

$$C_{xw} - \left[V_x\left(\frac{\partial A}{\partial B_x}\right) + C_{xw}\left(\frac{\partial A}{\partial B_w}\right)\right]\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\left[C'_{xw}\left(\frac{\partial A}{\partial B_x}\right) + V_w\left(\frac{\partial A}{\partial B_w}\right)\right]'. \tag{A12}$$

Appendix C: **Estimators of C and Σ**

This paper is of primarily concerned with estimation of the of the coefficients of the probit regression of Z on x and w. However, the overall estimator $\hat{\theta}$, as given by (A5), also provides efficient estimators of $\hat{C}$ and $\hat{\Sigma}$ of C and Σ Conniffe(1997) showed how estimation of a linear regression jointly with a probit employing the same explanatory variables, but with extra observations on the binary variable, leads to an improved estimator of the linear model. The $\hat{C}$ estimators from (A5) are the generalisation of this estimator to a set of linear equations – the l regressions of the w variables on the x variables. As regards $\hat{\Sigma}$, we are not interested in the components of Σ per se, except to the extent that some estimate is required to implement the asymptotically efficient estimators of B_x and B_w and their variances. Appendices A and B show that $\widetilde{\Sigma}$ suffices for that.

Appendix D: **The case of observed *Y***

When Y is observed the components of the likelihood are

$$L_{r,y|w} = -\frac{\mathrm{r}}{2}\log 2\pi - \frac{r}{2}\log\sigma_{yy.w} - \frac{1}{2\sigma_{yy.w}}\sum_{1}^{r}(y_i - x_i'B_x - w_i'B_w)^2,$$

$$L_{r,w} = -\frac{\mathrm{rl}}{2}\log 2\pi - \frac{r}{2}\log|\Sigma| - \frac{1}{2}\sum_{1}^{r}(w_i - C'x_i)'\Sigma^{-1}(w_i - C'x_i)$$

and

$$L_{n-r} = -\frac{\mathrm{n-r}}{2}\log 2\pi - \frac{n-r}{2}\log\sigma_{yy} - \frac{1}{2\sigma_{yy}}\sum_{r+1}^{n}(y_i - x_i'A)^2,$$

where

$$A = B_x + CB_w.$$

$\widetilde{B}_x$ and $\widetilde{B}_w$ are now the usual OLS estimators and V_x, V_w and C_{xw} are the corresponding variances and covariance, while $\widetilde{A}$ and $\overline{A}$ are OLS estimators of coefficients of *Y* on just the *x* variables for the *r* and *(n– r)* observations respectively. Then it is easily shown that (A9) become

$$\left(\frac{\partial A}{\partial B_x}\right)_{\tilde{\theta}} = I_k \text{ and } \left(\frac{\partial A}{\partial B_w}\right)_{\tilde{\theta}} = \widetilde{C}'$$

and (A7) and (A8) become

$$\hat{B}_x = \widetilde{B}_x - \left[\widetilde{V}_x + \widetilde{C}_{xw}\widetilde{C}'\right]\left(\overline{V}_{\overline{A}} + \widetilde{V}_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)$$

and

$$\hat{B}_w = \widetilde{B}_x - \left[\widetilde{C}'_{xw} + \widetilde{V}_w\widetilde{C}'\right]\left(\overline{V}_{\overline{A}} + \widetilde{V}_{\widetilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)$$

Remembering that $\sigma_{yy} = \sigma_{yy.w} + B_w'\Sigma B_w$,

$$\widetilde{V}_x + \widetilde{C}_{xw}\widetilde{C}' = \frac{\widetilde{\sigma}_{yy.w}}{\widetilde{\sigma}_{yy}}\widetilde{V}_{\widetilde{A}} = \frac{\widetilde{\sigma}_{yy.w}}{\widetilde{\sigma}_{yy.w} + \widetilde{B}_w'\widetilde{\Sigma}\widetilde{B}_w}\widetilde{V}_{\widetilde{A}},$$

where $\sigma_{yy.w}$ is simply estimated from the error mean square of regression of *Y* on the *x* and *w* variables for the *r* complete observations, and $\widetilde{C}_{xw} + \widetilde{V}_w \widetilde{C}' = 0$ then we obtain:

$$\hat{B}_x = \widetilde{B}_x - \frac{\widetilde{\sigma}_{yy.w}}{\widetilde{\sigma}_{yy}} \widetilde{V}_{\widetilde{A}} \left(\overline{V}_{\overline{A}} + \widetilde{V}_{\widetilde{A}} \right)^{-1} \left(\widetilde{A} - \overline{A} \right) \quad \text{and} \quad \hat{B}_w = \widetilde{B}_w .$$

These are the expressions given in (12) and (13) of the main text.

References

Allison, P. (2001): *Missing Data,* Thousand Oaks, CA; Sage Publications

Biancotti, C., G. D'Allessio and A. Neri (2008): "Measurement Error in the Bank of Italy's Survey of Household Income and Wealth," *Review of Income and Wealth*, 54(3), 466-492.

Brunello, G. and R. Miniaci (1999): "The Economic Returns to Schooling for Italian men. An Evaluation based on Instrumental Variables," *Labour Economics,* 6, 509-519.

Campbell , J. (2006): "Household Finance," *The Journal of Finance,* LXI(4), 1553-1604.

Cameron, C. and P. Trivedi (2005): *Microeconometrics: Methods and Applications*, Cambridge University Press.

Caseur, D. (2005): "Optimal Sampling from Concomitant Variables for Regression Problems," *Journal of Statistical Planning and Inference*, 128, 289-301.

Chen, Y. and H. Chen (2002): "A Unified Approach to Regression Analysis under Double-Sampling Designs," *Journal of the Royal Statistical Society B,* 62, 449-460.

Chesher, A. (1984): "Improving the Efficiency of Probit Estimators," *Review of Economics and Statistics*, 66, 523-527.

Cochran, W., (1963): *Sampling Techniques,* New York, Wiley.

Conniffe, D., (1983): "Small-Sample Properties of Estimators of Regression Coefficients given Common Pattern of Missing Data," *Review of Economic Studies*, 50, 111-120.

Conniffe, D., (1997): "Improving a Linear Regression through Joint Estimation with a Probit Model," *The Statistician*, 46, 487-493.

Conniffe, D. and M.A. Moran (1972): "Double Sampling with Regression in Comparative Studies of Carcass Composition," *Biometrics*, 28, 1011-1023.

Cox, D.R. and D.V. Hinkley (1974): *Theoretical Statistics*, London: Chapman and Hall.

Couch, K., (1992): "New Evidence on the long-Term Effects of employment Training Programs," *Journal of Labor Economics,* 10(4), 380-88.

Guiso, L., T. Jappelli and L. Pistaferri (2002): "An Empirical Analysis of Earnings and Employment Risk," *Journal of Business and Economic Statistics,* 20(2), 241-253.

Dagenais, M.G., (1973): "The Use of Incomplete Observations and Multiple Regression Analysis: a Generalised Least Squares Approach," *Journal of Econometrics*, 1, 317-328.

Dolton, P and D. O'Neill (1996): "Unemployment Duration and the Restart Effect: Some Experimental Evidence," *The Economic Journal,* cvi(435), 387-400.

Dolton, P. and D. O'Neill (2002): "The Long-Run Effects of Unemployment Monitoring and Work Search Programmes. Some Experimental Evidence from the U.K.," *The Journal of Labor Economics,* 20(2), 381-403.

Engel, B. and P. Walstra (1991): "Increasing Precision or Reducing Expense in Regression Experiments by using Information from Concomitant Variable," *Biometrics*, 47, 13-20.

Gourieroux, C. and A. Monfort (1981): "On the Problem of Missing Data in Linear models," *The Review of Economic Studies*, 48, 579-586.

Guiso, L., T. Japelli and D. Terlizzese (1996): "Income Risk, Borrowing Constraints and Portfolio Choice," *American Economic Review*, 86(1), 158-172.

Hamermesh, D., (1999): "LEEping into the Future of Labor Economics: The Research Potential of Linking Employer and Employee data," *Labour Economics*, 6(1), 25-41.

Hartog, J., A. Ferrer-i-Carbonell and N. Jonker (2002): "Linking Measured Risk-aversion to Individual Characteristics," *Kyklos*, 55, 3-26

Hausman, J.A., (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251-1271.

Heckman, J., (1976): "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economic and Social Measurement,* 5, 475-492.

Heckman, J., (1979): "Sample Selection bias as a Specification Error," *Econometrica*, 47, 153-161.

Horowitz, J., and C. Manski (2006): "Identification and Estimation of Statistical Functionals using Incomplete Data," *Journal of Econometrics,* 132, 445-459.

Jappelli, T., and L. Pistaferri (2006): "Intertemporal Choice and Consumption Mobility," *Journal of the European Economic Association*, 4(1), 75-115.

Kendall, M.G. and A. Stuart (1966): *The Advanced Theory of Statistics*, Volume 3, Griffin, London.

Little, R.J.A., (1993): "Regression with Missing Xs: A Review," *Journal of the American Statistical Association*, 87, 1227-1237.

Little, R. and D. Rubin (2002): *Statistical Analysis with Missing Data*, Wiley and Sons, New Jersey.

Paul, C., W.M. Mason, D. McCaffrey, and Sarah A. Fox (2008): "A Cautionary Case Study of Approaches to the Treatment of Missing Data," *Statistical Methods and Applications,* 17(3), 351-372.

Pistaferri, L., (2003): "Anticipated and Unanticipated Wage Changes, Wage Risk and Intertemporal Labour Supply," *Journal of Labour Economics*, 21(3), 729-754.

Rao, C. R., (1967): 'Least Squares Theory using an Estimated Dispersion Matrix and its Application to Measurement of Signals' in *Proceedings of 5th Berkley Symposium in Mathematical Statistics and Probability*, Vol. II, Berkley: University of California Press.

Robins, J., A. Rotnitzky and L. Zhao (1994): "Estimation of Regression Coefficients When Some Regressors are not Always Observed," *Journal of the American Statistical Association,* 89(427), 846-866.

Robins, J., M. Fushing Hsieh and W. Newey (1995): "Semiparametric Efficient Estimation of a Conditional Density with Missing or Mismeasured Covariates," *Journal of the Royal Statistical Society, Series B*, 57(2), 409-424.

Ronning, G. and M. Kukuk (1996): 'Efficient Estimation of Ordered Probit Models', *Journal of the American Statistical Association* 91, 1120-1129.

Rosen, H. and S. Wu (2004): "Portfolio choice and Health Status," *Journal of Financial Economics,* 72, 457-484.

Royston, P. (2004): "Multiple Imputation of Missing Values," *The Stata Journal*, 4(3), 227-241.

Rubin, D.B., (1974): "Characterising the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467-474.

Rubin, D.B., (1976): "Inference and Missing Data," *Biometrika*, 63, 581-592.

Schafer, J. L., (1997): *Analysis of Incomplete Multivariate Data*, London, Chapman and Hall.

Seber, G., (2008): *A Matrix Handbook for Statisticians,* Wiley and sons, New Jersey.

Wooldridge, J.M., (2006): *Introductory Econometrics – A Modern Approach,* ISE (Mason OH, Thompson South-Western)

Table 1:
Monte Carlo Study: Comparison of the Efficient Estimator with the Complete Case Probit Estimator and Multiple Imputation approaches

N=500	Complete Case Analysis				Efficient Estimator				
%missing	$\tilde{B}_w$	$\tilde{B}_x$	$\tilde{V}_w$	$\tilde{V}_x$	$\hat{B}_w$	$\hat{B}_x$	$Var(\hat{B}_w)$	$Var(\hat{B}_x)$	$\frac{Var(\hat{B}_x)}{\tilde{V}_x}$
10% Imputation	1.019	1.016	.015	.023	1.02 .99	1.006 1.008	.015 .016	.021 .021	.93
25% Imputation	1.026	1.020	.018	.028	1.026 .95	1.007 .98	.018 .019	.022 .022	.81
50% Imputation	1.038	1.032	.028	.042	1.04 .89	1.011 .96	.028 .028	.025 .024	.60
70% Imputation	1.072	1.049	.051	.076	1.075 .85	1.016 .94	.051 .038	.033 .027	.43
N=1000									
10% Imputation	1.013	1.01	.007	.011	1.013 .98	1.006 .99	.007 .008	.010 .01	.93
25% Imputation	1.015	1.011	.0087	.0133	1.015 .94	1.006 .98	.0087 .009	.0108 .01	.80
50% Imputation	1.022	1.014	.013	.0203	1.022 .87	1.006 .95	.013 .013	.012 .011	.60
70% Imputation	1.032	1.026	.023	.035	1.03 .84	1.01 .94	.0228 .019	.0145 .013	.41
N=6000									
10% Imputation	1.002	1.003	.0012	.0018	1.002 .97	1.002 .99	.0012 .0012	.0017 .0017	.92
25% Imputation	1.003	1.003	.0014	.0022	1.003 .93	1.002 .96	.0014 .0015	.00175 .0017	.81
50% Imputation	1.004	1.004	.0021	.0033	1.004 .87	1.002 .94	.0021 .002	.00195 .0018	.60
70% Imputation	1.007	1.003	.0035	.0055	1.007 .83	1.002 .92	.0035 .003	.0023 .0021	.42

Table 2
Summary Statistics

Variable Name	Complete Sample	Subsample
Risky assets	23.5%	30.2%
Age	54	51.2
College Education	10.3%	11.6%
Gender	31.5%	29.35%
Married	71.3%	74%
Region 1 – North-East	27.2%	28%
Region 2 – North-West	22.5%	26.9%
Region 3 – Centre	22.1%	15.7%
Region 4 – South	18.6%	20.9%
Region 5 – Islands	9.6%	8.45%
Risk Aversion		.1778
Sample Size	6779	1029

Table 3
Determinants of Portfolio Allocation among Italian Households.
Dependent Variable is a Binary Variable taking the value 1 if Respondents are Identified as having Held Risky Assets at the end of 2000.

Independent Variable	Coefficient	Standard Error	Coefficient	Standard Error
	Complete Case analysis		Efficient Estimator	
Constant	-1.24	.55	-.96	.27
Age	.06	.02	.04	.01
Age-Squared	-.0006	.0002	-.0004	.0001
College	.65	.13	.62	.06
Gender	.01	.10	-.09	.05
Marital Status	.18	.11	.21	.05
North-West	.15	.11	.18	.05
Centre	-.22	.13	-.21	.05
South	-.50	.13	-.72	.06
Islands	-.98	.21	-.70	.08
Risk-Aversion	-4.08	.77	-3.9	.76