IZA DP No. 4081

Efficient Probit Estimation with Partially Missing Covariates

Denis Conniffe
Donal O'Neill

March 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Efficient Probit Estimation with Partially Missing Covariates

Denis Conniffe
National University of Ireland Maynooth

Donal O'Neill
*National University of Ireland Maynooth
and IZA*

Discussion Paper No. 4081
March 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4081
March 2009

ABSTRACT

Efficient Probit Estimation with Partially Missing Covariates[*]

A common approach to dealing with missing data is to estimate the model on the common subset of data, by necessity throwing away potentially useful data. We derive a new probit type estimator for models with missing covariate data where the dependent variable is binary. For the benchmark case of conditional multinormality we show that our estimator is efficient and provide exact formulae for its asymptotic variance. Simulation results show that our estimator outperforms popular alternatives and is robust to departures from the benchmark case. We illustrate our estimator by examining the portfolio allocation decision of Italian households.

JEL Classification: C25, G11

Keywords: missing data, probit model, portfolio allocation, risk aversion

Corresponding author:

Donal O'Neill
Economics Dept.
NUI Maynooth
Maynooth, Co. Kildare
Ireland
E-mail: donal.oneill@nuim.ie

[*] We would like to thank Greg Conor, Olive Sweetman and seminar participants at the NUI Maynooth Economics seminar series and the 2008 Irish Economic Association annual conference for helpful comments on an earlier draft of this paper.

1. Introduction

Many approaches for dealing with missing data in regression type analyses have appeared in both the econometrics and mainstream statistical literature. Reviews of the statistical literature 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 include Horowitz and Manski (2006) and Wooldridge (2009), with an overview provided in Cameron and Trivedi (2005). Yet enthusiasm for the practical application of the methods seems muted at best. For example, the popular econometrics textbook by Wooldridge (2009, p. 322) notes that while missing data is common in real world applications, the improvement from using alternative estimators "is usually slight, while the methods are somewhat complicated. In most cases, we just ignore the observations that have missing information."

Both the practical complications and the lack of efficiency gains cited by Wooldridge are most easily overcome when parametric estimation, tailored to the model of interest, is applied to a tractable data pattern. This paper estimates a probit equation when some explanatory variables are unrecorded on r of the original n observations, but the binary dependent variable and the remaining explanatory variables are recorded on all observations. We show that explicit formulae for coefficient estimators and their variances are quite straightforward and easily implemented on standard econometrics packages such as Stata. Taking multinormality as an initial benchmark case we show the estimator is efficient and that improvements over complete case analysis can be very substantial. We show by simulation and analysis of real data, that our estimator can outperform other popular techniques for dealing with missing data. By considering departures from the initial model we also show that these findings extend beyond the benchmark case.

Any approach to missing data analysis requires assumptions about the process causing the absence of data. There is a large literature on this topic with Rubin (1974, 1976) of central importance. We assume data are missing at random (MAR), that is, that the probability of data being missing on *W,* say, is unrelated to the value of *W* conditional on other variables in the model. We will discuss the plausibility of MAR and show how to test for it later.[1] In later sections we also discuss testing of the parametric assumptions embodied in the demonstration of efficiency, assess robustness of our estimator to departures from these assumptions and consider what modifications to our estimator, if any, may be appropriate in these circumstances.

Our ability to obtain closed form expressions for the estimator and its variance is facilitated by the tractable missing data pattern we consider. Fortunately this pattern is quite common in real world data sets and often arises when collecting data on all variables, from all respondents, is expensive or otherwise difficult. In this case deliberate "double sampling" for surveys, as described by Cochran (1963), is often used. A large scale example of such a procedure was that adopted by the U.S. Bureau of the Census when collecting the 2000 Census data. Each household received either a short-form or a long-form. The long-form questionnaire included the same 6 population questions (related to age, gender and marital status) and 1 housing question as on the Census short-form, *plus* 26 additional population questions (including education, health, employment status and income) and 20 additional housing questions. About 1 in every 6 households received the long form, giving rise to this paper's data structure. This pattern, generated by deliberate double sampling, can be reinforced when researchers try to match Census data across different years.[2]

The data structure of r complete and $(n–r)$ incomplete observations also arises frequently in econometrics through mechanisms other than deliberate random sampling. In many fields, such as labour economics, there is a growing tendency to draw data from multiple sources. Often the sample sizes can differ between two data sources. Dolton and

[1] If the process generating the missing data is (using Rubin's term) non-ignorable, inference based on the complete observations alone may not be representative of the population of interest. Correct inference may be obtainable by joint modelling of the process along with the model of substantive interest, although this requires extra assumptions as, for example, with Heckman's (1976, 1979) sample-selection models. For the worst case scenario of no prior information Horowitz and Manski (2006) propose bounds for parameter magnitudes, but applications often find the resulting intervals too wide to be useful.

[2] Beenstock (2004) estimates an income mobility model using matched Israeli Census data in which complete case analysis uses only 20% of the base sample.

O'Neill (1996) evaluated a government training programme in the UK where data on personal characteristics such as sex, age and treatment status, along with some outcome data, were obtained at the initial interview stage for a sample of 8925 individuals. 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, but completed by only 5200 of the original sample.

Even when there is no timing difference in the two data sources, one source may be more prone to non response than the other. In using linked employer-employee data sets for example (for a review see for Hamermesh (1999)) firm related data such as tenure, wages and firm size are often available for all respondents from payroll data, whereas individual level data such as education and health require individual surveys. Differences in the response rate across firms and workers can give rise to our data structure. This situation also arises when combining administrative and survey data, where the administrative data provide measures such as earnings or unemployment histories, with limited personal data (often age and gender) and the survey data are used for more detailed personal characteristics such as education, marital status and family size (examples include the long-run evaluations of training programmes by Couch (1992) and Dolton and O'Neill (2002)). Researchers in this situation have either used the full sample restricted to the subset of variables obtained from the administrative data or the full range of explanatory variables for the complete cases only. Neither approach is ideal.

Thus while a range of missing data patterns can occur in practice, these examples show that our assumed pattern of r complete and $(n–r)$ incomplete observations is not only tractable, but also relevant in real world settings.

Our application examines the portfolio allocation decisions of Italian households using the Bank of Italy's Survey of Household Income and Wealth (SHIW). A major advantage of SHIW for the study of portfolio allocation is that it contains a question that permits 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 most of the missing data is ignorable by design and where complete case analysis involves dispensing with over half of the original sample. Using our estimator on all the data reduces the estimated standard errors of the coefficients greatly compared to complete case analysis and several coefficients, previously imprecisely estimated, become significant. Such dramatic changes are a clear illustration of the potential gains which may be achieved by using all available data in an efficient manner.

The structure of the paper is as follows. Section 2 specifies the model. Section 3 presents the efficient estimator for this model and obtains explicit formulae for the asymptotic variance of our estimator. Section 4 describes how to test the assumptions underlying our estimator and discusses modifications to our estimators that might be required as a consequence. Section 5 briefly discusses extensions of our approach to some other models. 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

This paper focuses on estimation of a probit binary choice model. We follow standard practice and relate the observed binary variable Z_{i} to an underlying unobserved continuous latent variable Y_{i}, as follows :

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

As in standard probit analysis, we assume a linear index function

$$Y_i = X_i^{'}B_x + W_i^{'}B_w + \varepsilon_i \tag{2}$$

where X_i and W_i are (k x 1) and (l x 1) vectors of regressors, independent of ε_i, which is assumed to be distributed N(0,1).[3] We consider situations where data are available on $\{X_{i,}, W_{i}, Z_i\}$ for i=1….r. This represents the complete observation sample. In addition

[3] The choice of a unit variance matches the conventional assumption of standard probit analysis.

there are the further (*n-r*) observations on which {$X_{i,}, Z_i$} alone are measured. To utilise these additional observations we initially assume

$$W_i^{'} = X_i^{'}C + u_i^{'} \qquad (3)$$

where C is a (k x l) matrix of parameters, $u_i^{'} \sim MVN(0,\Sigma)$. These assumptions are both convenient analytically and permit efficient estimation. We show later that these assumptions are testable and in section 4 we discuss possible approaches in the event that the parametric assumptions are rejected. However, the results presented in section 6 indicate that the desirable properties of our estimator are robust to many plausible departures from the parametric assumptions in (3).

In conjunction with (2), equation (3) implies that, conditionally on X only, $Y_i = X_i^{'}(B_x + CB_w) + e_{yi}$ with variance $\sigma_{yy} = 1 + B_w^{'}\Sigma B_w$ and $(e_{yi}, W_i^{'})$ are multivariate normally distributed. Sometimes a stronger assumption, namely that *all* regressors {$X_{i,}$, W_i} are joint normally distributed, is made in the literature on discrete choice analysis (see Greene, 2008, p. 810), accompanied by a warning that occurrence of dummy variables etc. would invalidate this assumption. Dummy variables that appear in the {X_i} variables require no special attention or assumptions in our model, while the case where dummy variables appear in {W_i} is considered in section 6.

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. 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 Estimators and Variances

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

$$\hat{\theta} = \overset{\approx}{\theta} + J(\overset{\approx}{\theta})\frac{\partial}{\partial\theta}L_n(\overset{\approx}{\theta}) \qquad (4)$$

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).

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. $\tilde{B}_x$ and $\tilde{B}_w$ are the coefficients from a standard probit analysis with X and W as explanatory variables, $\tilde{C}$ is $(\tilde{c}_1, \tilde{c}_2, .\, \tilde{c}_l)$, where $\tilde{c}_j$ is the OLS coefficient vector for regression of the jth W on the X variables and $\tilde{\Sigma}$ is the estimator of Σ based on the OLS residuals. As $\tilde{\theta}$ 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:

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

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

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

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:

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

and

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

where $A = \dfrac{(B_x + CB_w)}{\sqrt{1 + B_w^{'}\Sigma B_w}} = \dfrac{1}{\sqrt{\sigma_{yy}}}(B_x + CB_w)$ and $\widetilde{A}$ results from A by replacing θ by $\widetilde{\theta}$. Note that $\widetilde{A}$ is *not* obtained from the simple probit of Z_i on X_i for the *r* complete observations, which we will denote as A^*, but makes use of the multinormality assumption on the *W*s to improve on that simple probit. In fact $\widetilde{A}$ is actually Chesher's (1984) efficient estimator for joint estimation of a probit equation and a system of linear equations with the same explanatory variables. $\overline{A}$ is the probit estimator obtained from the *(n-r)* incomplete observations where the underlying index function is $Y_i = X_i^{'}(B_X + CB_W) + e_{yi}$. $\tilde{V}_{\tilde{A}}$ and $\bar{V}_{\bar{A}}$ are the corresponding 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 from $L_{r,z|w}$ and $\tilde{C}_{xw}$ their estimated covariance matrix.

The asymptotic variances of $\hat{B}_x$ and $\hat{B}_w$ are derived in Appendix B. Since $\hat{B}_x$ and $\widetilde{B}_x - \left[V_x\left(\dfrac{\partial A}{\partial B_x}\right) + C_{xw}\left(\dfrac{\partial A}{\partial B_w}\right)\right]\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\left(\widetilde{A} - \overline{A}\right)$ differ only in terms of $O_p(n^{-1})$ they have the same asymptotic variance so

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

and the estimated variance is obtained by replacing the Vs by $\widetilde{V}s$, C_{xw} by $\tilde{C}_{xw}$ and the derivatives by their values evaluated at $\tilde{\theta}$. Similarly, the variance of $\hat{B}_w$ may be shown to be

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

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

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

While it is intuitively obvious from the data structure that the variance of $\hat{B}_x$ may be much smaller than that of $\widetilde{B}_x$, we would not expect the same for the variance of $\hat{B}_w$ relative to $\widetilde{B}_w$. This is because we have extra observations on the X variables from the (n – r) incomplete observations, but no extra information on the W variables. We will discuss this in more detail later in the paper.

4. Testing Assumptions and Possible Modifications

From examination of (7) and (8) it is clear that consistency of $\hat{B}_x$ and $\hat{B}_w$ requires that $(\tilde{A} - \overline{A})$ be a consistent estimator of zero. A necessary condition for this is that the missing data for W are MAR. The MAR assumption is obviously valid if the r observations have been deliberately chosen at random from the n. Otherwise the assumption can be assessed by comparing the coefficient estimates based on complete data with the estimates using all data. If true, the former are consistent, but inefficient, while the latter are consistent and efficient. If the estimates look quite similar with reduced standard errors for the estimates based on all the data the assumption is probably true. If the estimates look very different the assumption is probably untrue. More formally, a Hausman (1978) type test can be performed based on the explicit variance formulae derived in the previous section. The efficient estimator is $\hat{B}_x$ with variance given by (9) and $\widetilde{B}_x$ is a consistent estimator with variance V_x. The asymptotic variance of the difference between an efficient estimator and another consistent one is the difference of the variances and so

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

is asymptotically χ^2 with k degrees of freedom.

If the MAR assumption is inappropriate then the above χ^2 test should prove statistically significant. However it is also possible that the test may prove significant when the MAR assumption is true but another maintained assumption of our estimator is false. For instance in deriving our estimator we estimated A by $\widetilde{A}$. If the normality assumption underlying Chesher's estimator is invalid, $\left(\tilde{A}-\bar{A}\right)$ might not have (asymptotic) expectation zero or at least, $\widetilde{A}$ might not be the efficient estimator of A under the true (unknown) joint distribution. Correspondingly $\hat{B}_x$ would no longer be efficient. However, we can investigate this assumption by comparing $\widetilde{A}$ and $A\,*$, again employing a variant of the Hausman test, this time using only the r complete observations, the estimate (A3) of the variance of $\widetilde{A}$ and the standard probit formula for the variance of $A\,*$.

If MAR seems appropriate, but joint normality is not, several approaches are possible. The device, going back at least to Rao (1967), of modifying a consistent estimator $\widetilde{\theta}$ through

$$\hat{\theta} = \widetilde{\theta} + \Omega\, S, \tag{12}$$

where S is a statistic correlated with $\tilde{\theta}$, with asymptotic expectation zero and Ω is a constant could be employed. We could choose $\widetilde{\theta} = \widetilde{B}_x$ and $S = A\,* - \bar{A}$. The resulting estimator resembles (7) with $A\,*$ instead of $\widetilde{A}$, but may have some disadvantages. First as demonstrated by Chesher (1984) in SURE estimation of a probit and by Ronning and Kukuk (1996) for the ordered probit problem, failure to exploit joint normality when it does hold can imply substantial loss of estimation efficiency, suggesting that these estimators should not be set aside lightly. Second, not every value of Ω will achieve a significant, or perhaps any, improvement over $\tilde{B}_x$ and it is unclear how to choose Ω without assumptions. For minimum variance in its class Ω should be the covariance of $\widetilde{B}_x$ and $A^* - \bar{A}$ multiplied by the inverse of the variance of $A^* - \bar{A}$. But $\bar{A}$ and A^* are functions of the Z and W variables and without our distributional assumptions this optimal Ω cannot be calculated. A possible way forward is to remember that $\hat{\theta}$ is consistent for any constant Ω, even if inefficient, and to use the multinormal case multiplier derived for $\tilde{A} - \overline{A}$

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

and its corresponding estimate, as in (7).

Another approach is to continue to use the estimators outlined in (7) even when the joint normality assumption is untrue in the hope that the estimator is reasonably robust. As will be seen in Section 6, this option works well for many of the examined departures from the benchmark models, some of which involve extreme departures from normality.

5. Other Models

While this paper is primarily concerned with estimation of the coefficients of a probit regression of *Z* on *X* and *W*, the estimator $\hat{\theta}$, as given by (A5) in Appendix A, provides solutions to other models that might well be relevant for our data structure. If an equation of interest relates a continuous dependent variable, *W*, measured only on the *r* observations, to a set of explanatory *X* variables, it is well known that extra observations on just the explanatory variables cannot increase the precision of estimation. But joint estimation with another dependent variable, measured on all observations can do so. When that variable is binary and modelled by a probit, the efficient estimator is that given by Conniffe (1997). The $\hat{C}$ estimators from (A5) are the generalisation of that estimator to a set of *l* linear equations of the *W* variables on the *X* variables. The overall model may be viewed as seemingly unrelated regressions with one dependent variable binary, recorded on the extra $n - r$ observations.

Another interesting model arises if the dependent variable *Y* is continuous and observed for all observations and we want to estimate its regression on the *W* variables and the *X* variables. In Appendix C we show that, with the same joint normality assumptions as in Section 2, efficient estimators of the regression coefficients on X and W for this linear model are

$$\hat{B}_x = \tilde{B}_x - \frac{\tilde{\sigma}_{yy.w}}{\tilde{\sigma}_{yy}} V_{\tilde{A}} \left(V_{\bar{A}} + V_{\tilde{A}} \right)^{-1} \left(\tilde{A} - \bar{A} \right) \text{ and } \hat{B}_w = \tilde{B}_w \tag{13}$$

respectively, where $\tilde{B}_x$ and $\tilde{B}_w$ are the usual OLS estimators, $\tilde{A}$ and $\bar{A}$ are OLS estimators of coefficients of Y on just the X variables for the r and $(n-r)$ observations, $\sigma_{yy.w}$ 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. Failure to improve on $\tilde{B}_w$ is intuitively plausible for the reason mentioned at the end of Section 3.[4] These estimators are not new and, written a little differently, were obtained previously by Conniffe (1983). Conniffe's approach was not based on the likelihood approximation used in this paper but rather on the device (12) as discussed above. In this linear case $\hat{C}$ provides the seemingly unrelated regression estimators for the system with Y and the W variables as regressands and the X variables as regressors (see also Conniffe (1985)).

Models for seemingly unrelated regression with extra observations for an equation are closely related to models that have appeared in the statistical literature on the use of auxiliary or surrogate outcome data. Some authors, for example, Pepe (1992), Pepe, Reilly and Fleming (1994) and Chen and Chen (2002) have employed partially semi-parametric approaches that permit relaxation of the assumptions about joint distributions. However, the estimators often cannot be implemented without imposing strong and possibly implausible conditions on data, and even then can involve substantial loss of information compared to parametric analysis.[5]

[4] The ML estimator of B_w differ from $\tilde{B}_w$ but its variance is asymptotically the same.

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 first simulations is

$$Y_i = X_i^{'}B_x + W_i^{'}B_w + \varepsilon_i \tag{14}$$

where *X* and *W* are both scalar random variables and ε_i ~N(0,1). For the simulation we assume that $X_i^{'} \sim N(0,1)$. In addition:

$$W_i^{'} = X_i^{'}C + u_i^{'} \tag{15}$$

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

$$\begin{aligned} Z_i &= 1 \quad \text{if } Y_i > 0 \\ Z_i &= 0 \quad \text{if } Y_i \le 0 \end{aligned} \tag{16}$$

We consider situations where data are 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 at random. For most of the simulations presented in the paper we choose 3 missing mechanisms that generate approximately 25%, 50% and 75% missing data respectively. The precise missing mechanisms are Pr(M=1|X_i)=Φ(X_i-1), Pr(M=1|X_i)=Φ(X_i) and Pr(M=1|X_i)=Φ(X_i+1), respectively, where Φ(.) denotes the cumulative standard normal distribution function.[6] In this paper we present the simulation results for *n*=1000, though we have carried out the analysis with other sample sizes with very similar results.

[5] This topic has been discussed in detail in Conniffe (1996). Chen and Chen (2002), using the idea represented by (12), do derive a partially semi-parametric estimator for linear seemingly unrelated regressions equivalent to that of Conniffe (1985).

[6] Similar missing mechanisms were used in Little and Rubin (2002) to illustrate the MAR assumption.

The results of the simulations, are given in Table 1. The estimates for our new estimator are easily obtained from a new user-written Stata package provided by the authors.[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 both estimators that goes to zero as $r \rightarrow \infty$.[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 simulations.

However, when we turn to the estimated variances we see significant improvements in precision when our efficient estimator is used. 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 $\tilde{B}_x$ shows 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 75% of the data are missing, we see an approximate seventy 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 the order of twenty percent when we consider missing data of the order of twenty five percent of the initial sample.[9]

The results presented in Table 2 allow us to compare the performance of our estimator to other popular approaches used with missing data. For ease of exposition we repeat the results from our efficient estimator in the first row of each panel. Underneath the

[7] 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.

results for our efficient estimator we provide two sets of estimates based on imputation techniques. The first row of results uses mean imputation, where the missing values for *W* are replaced by the mean of the observed *W*. This variable is used instead of the observed *W* in the probit, along with a dummy variable indicating whether or not the observation was imputed. In addition to simple mean imputation we also report estimates based on a popular multiple imputation technique for handling missing data. Underneath the estimates based on mean imputation, we present 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 *X*, conditional on the observed values and the variables in $\{Z,X\}$.[10]

The results reported in Table 2 indicate biases in both imputation techniques. For mean imputation, the estimated coefficient on *X* is biased upwards, with the bias being large even for moderate degrees of missing data. Although still widely used in practice, these results support earlier claims that simple mean imputation is not a satisfactory way of dealing with missing data.[11] On the other hand we see that the coefficients on both the complete and partially observed variables are biased towards zero with the multiple imputation approach, with the biases growing as the proportion of missing data increases.[12]

As we noted earlier the assumption of conditional normality of the missing regressors, W_i, permits efficient estimation of our model. In the remainder of this section we use simulation methods to examine the robustness of our estimator when the normality assumption fails. The first departure from normality is a rather mild one whereby we assume the *W* variable has a logistic rather than a normal distribution. This allows us to examine the

[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] We also compared the performance of our estimator to the inverse probability weighted estimator discussed by Wooldridge (2007). However, given our structure this latter estimator is dominated by the unweighted complete case analysis and so the results are not presented (see Wooldridge 1999).

[11] For a related discussion in the context of the linear regression model see Jones (1996).

[12] 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

robustness of our estimator to heavy tailed distributions. In a second case we maintain the continuity assumption but allow the regressor to be uniformly distributed. The third model takes a more dramatic departure from normality and considers the case where the missing regressor is a binary variable dependent on the observed regressors.

The results of the simulations for each of these three cases are given in Table 3. These results show that our estimator is very robust to these departures from conditional normality in the missing regressors. With both the logistic and uniform models the estimated efficiency gains are similar to the normal case. More surprising perhaps is the fact that we observe larger efficiency gains when our estimator is applied to a missing binary regressor than we did with the continuous normal regressor. Interestingly, and unlike the continuous regressor case, the simulation results presented in the fourth row of Table 3 show that the multiple imputation approach also performs well in the case of missing binary data. Further simulations showed that our estimator achieved efficiency levels close to those obtained by the MLE with no missing data. The exceptional performance of both estimators in this situation suggests that missing data on a binary regressor can be effectively dealt with using either approach. The intuition for this result is easiest to see in the multiple imputation approach. When the missing regressor is binary imputation need only impute the sign of the underlying latent variable in order to assign it a zero or one. In this case small errors in the imputation of the level of the underlying latent variable for the missing regressor, that do not affect its sign, will not affect the final estimate. In this sense imputation of a twofold classification of a missing variable is more forgiving than imputation of a continuous variable, in turn leading to greater efficiency gains. While our estimator involves no imputation, the efficiency gains nonetheless derive from the extra (n-r) observations on the observed X and a similar intuition can be applied.

In the final part of this section we consider the possibility that equation (3) which relates W to X is miss-specified, either in the sense that a variable, G, has been omitted from

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.

the model for *W* or that the parametric relationship between *W* and *X* has been miss-specified. In the simulations here we consider cases where the true relationship between *W* and *X* is quadratic, though similar reasoning carries over to other cases.

The results in the top panel of Table 4 show that the general case of omitted variables in the model for *W* is not a problem for our estimator. Our estimator remains consistent and provides large efficiency gains over complete case analysis even when the omitted variable *G* is correlated with *X* (a correlation of .5 was chosen for the simulation presented). These results follow from the assumption of MAR.

The middle panel of Table 4 however, shows that failure to account for nonlinearity in the *W* function causes a problem for our estimator. In these simulations, where we assume that the true model for *W* is quadratic in *X* but fail to account for it, our estimator for B_x is no longer consistent. Vitally however, we find that our Hausman test is able to identify this problem and moves us to investigate misspecification further. The nature of the problem is that non-linearity of *W* in terms of *X* implies non-linearity of the marginal distribution of *Y* on *X*. However, estimation of $\tilde{A}$ in our approach assumes a linear marginal model and is thus not a consistent estimator for *A*.

One possible approach in this case is to use $(A^* - \overline{A})$ in our adjustment. Given MCAR $(A^* - \overline{A})$ has asymptotic expectation zero, even though neither estimator is a consistent estimator of the true coefficient. However, this approach no longer works if we consider the weaker assumption of MAR. $(A^* - \overline{A})$ need no longer have expectation zero when the distribution of *X* differs in the *r* complete and *(n-r)* incomplete samples. This is most easily seen in the linear regression model. In that case the standard formulae for omitted variable bias when *Y* is regressed on *X*, omitting X^2, will involve the estimated coefficient from the regression of *X* on X^2. This in turn will be a function of the moments of the distribution of *X*, which need not be equal across the complete and incomplete samples, even when MAR is assumed to be true.

We briefly discuss two ways to proceed in this instance. In cases where the Hausman test indicates a problem of misspecification we could simply adjust the models for both *Y* and *W* to include non-linear terms until the test no longer rejects. Although this may be appropriate when the number of missing is very large it is actually inferior to complete case analysis otherwise. The reason is that estimating a coefficient on X^2 that is zero in the true model of *Y* given *W* ignores the information in a zero-restriction. This is not accounted for in our adjusted estimator but is automatically imposed in the complete case analysis. Nevertheless, this approach may still be useful for a number of reasons. Firstly it can help in establishing the order of adjustment needed for the functional form of *W*. As noted earlier $\left(\tilde{A}-\overline{A}\right)$ is unlikely to have mean zero if the functional form of W is miss-specified. This suggests experimenting with expansions of *W* until the estimates of $\widetilde{A}$ are sufficiently close to $\overline{A}$.[13] Once we have achieved a suitable expansion the estimate of $\left(\tilde{A}-\overline{A}\right)$ from this extended model can be used to obtain the mean zero component of our required adjustment. To make use of the information in the zero-restriction we propose using this adjustment in conjunction with the original $\tilde{B}_x$ to form our new estimator. The results in the bottom panel of Table 4 show that this simple adjustment to our original estimator still leads to substantial efficiency gains over complete case analysis even if no further adjustment is made to our weighting matrix.

7. Empirical Application to Portfolio Allocation.

Campbell (2006) presents an overview of recent theoretical[14] and empirical[15] 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

[13] In our simulations this was achieved with a quadratic in *X* as expected.

[14] More detailed discussion of the theory underlying household portfolio decision making is provided by Gollier (2002).

allocation decisions of Italian households using the Bank of Italy's Survey of Household Income and Wealth (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 illustrates 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.[16]

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

[15] Previous empirical studies of portfolio allocation among households include Feldstein (1976), Guiso et al (1996), Bertaut and Starr-McCluer (2002), and Rosen and Wu (2004).

[16] 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[17]:

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

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 estimating this model would be to focus on the complete data. However in our application

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

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 5 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.[18] 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 6 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 the probability of

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

holding risky assets is highest among the middle-aged and more highly educated.[19] Those located in the south or the islands are less likely to hold risky assets.[20] Of the remaining coefficients neither the gender, marital status nor 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. Applying the Hausman test described in Section 4 gives a test statistic, which under the assumptions of MAR and joint normality, is asymptotically χ^2 with 10 degrees of freedom. The resulting value is 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 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.

[19] These results are consistent with previous studies of portfolio allocation (e.g Guiso et al (1996), Bertaut and Starr-McCluer (2002) and Rosen and Wu (2004)), though these studies had no or only limited controls for individual risk-aversion.

[20] The omitted region refers to those living in the North-East.

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 for a benchmark model and relate these to other approaches discussed in the literature. We also carry out simulations which illustrate that our estimator outperforms popular alternative approaches and also show that the performance of the estimator is robust to many departures from the benchmark case.

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 verify the validity of the assumptions underlying our estimator. 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, its robustness to distributional assumptions and the ease with which the estimator can be implemented with standard software packages 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 $\tilde{B}_x$ and $\tilde{B}_w$ are independent of $\tilde{C}$ and $\tilde{\Sigma}$.

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 $\tilde{A}$ results from A by replacing θ by $\tilde{\theta}$.

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

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

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

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

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

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

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

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

.

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

(A3) $$\tilde V_{\tilde A} = \left(\frac{\partial A}{\partial\theta}\right)'_{\tilde\theta}\tilde V_{\tilde\theta}\left(\frac{\partial A}{\partial\theta}\right)_{\tilde\theta}$$

and so

(A4)

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

Using (A2) and (A4)

`

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

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

Using this the efficient estimator, $\breve\theta = \tilde\theta + J(\tilde\theta)\dfrac{\partial}{\partial\theta}L_n(\tilde\theta)$, can be written as

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

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

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

is an efficient estimator. Denoting the variance matrix of $\tilde{B}_x$ by V_x, that of $\tilde{B}_w$ 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 $vec\,\tilde{C}$, 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).

$\tilde{V}_{\tilde{\theta}}$, the estimator of $V_{\tilde{\theta}}$, is obtained by replacing V_x, V_w and C_{xw} in (A6) by $\tilde{V}_x, \tilde{V}_w$ and $\tilde{C}_{xw}$ respectively, where these are produced by the standard probit regression for the r complete observations, and Σ and H by $\tilde{\Sigma}$ and $\tilde{H}$, where the σ_{ij} are replaced by their estimators based on OLS residuals. From the structure of (A6) it is clear that

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

and

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

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

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

where $\tilde{\sigma}_{yy} = 1 + \tilde{B}_w^{'}\tilde{\Sigma}\tilde{B}_w$. 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

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

and

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

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

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

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 $\tilde{B}_x$ and $\bar{A}$ are independent, the (asymptotic) variance of $\hat{B}_x$ is

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

$$= V_x - \mathrm{cov}(\tilde{B}_x, \tilde{A})\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\Lambda' - \Lambda\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\mathrm{cov}(\tilde{A}, \tilde{B}_x) + \Lambda\left(V_{\bar{A}} + V_{\tilde{A}}\right)^{-1}\Lambda'$$

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

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

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

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

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

This is equation (9) in the main text.

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

(A11)

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

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

(A12)

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

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

When *Y* is observed the components of the likelihood are

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

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

and

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

where

$$A = B_x + CB_w .$$

$\tilde{B}_x$ and $\tilde{B}_w$ are now the usual OLS estimators and V_x, V_w and C_{xw} are the corresponding variances and covariance, while $\tilde{A}$ and $\bar{A}$ 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

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

and (A7) and (A8) become

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

and

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

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

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

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 $\tilde{C}_{xw} + \tilde{V}_w\tilde{C}' = 0$ then we obtain:

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

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

References

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

Beenstock, M. (2004): "Rank and Quantity Mobility in the Empirical Dynamics of Inequality," *Review of Income and Wealth*, Series 50, no. 4, pp. 519-541.

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.

Bertaut, C. and M. Starr-McCluer (2002): "Household Portfolios in the United States," in L. Guiso, M. Haliassos and T. Jappelli eds.: *Household Portfolios*, MIT Press, Cambridge, MA.

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.

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. (1985): "Estimating regression equations with common explanatory variables but unequal numbers of observations", *Journal of Econometrics* 27, 179-196.

Conniffe, D (1996) “A comment on the use of auxiliary or surrogate outcome data”, *Journal of Statistical Planning and Inference 55*, 353-359.

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

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.

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.

Feldstein, M. (1976): “Personal Taxation and Portfolio Composition: An Econometric Analysis,” *Econometrica*, 44(4), 631-650.

Gollier, C. (2002): “What does Classical Theory Have to Say About Household Portfolios?” in L. Guiso, M. Haliassos and T. Jappelli eds.: *Household Portfolios*, MIT Press, Cambridge, MA.

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

Greene, W. H. (2008): *Econometric Analysis* 6th Edition, Pearson, New Jersey.

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

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.

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.

Jones, M.P. (1996): “Indicator and stratification methods for missing explanatory variables in multiple linear regression,” *Journal of the American Statistical Association,* 91,222-230

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.

Pepe, M. S. (1992): “Inference Using Surrogate Outcome Data and a Validation Sample”, *Biometrika* 79, 355-365.

Pepe, M.S., M. Reilly and T.R. Fleming (1994): “Auxiliary Outcome Data and the Mean Score Method”, *Journal of Statistical Planning and Inference,* 42, 137-160.

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.

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., (1999): “Asymptotic Properties of Weighted M-Estimators for Variable Probability samples,” *Econometrica***,** 67(6), 1385-1406.

Wooldridge, J.M., (2007): “Inverse Probability Weighted Estimation for General Missing Data Problems,” *Journal of Econometrics*, 141, 1281-1301.

Wooldridge, J.M., (2009): *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 under MAR assumption

	Complete Case Analysis				Efficient Estimator				
Approx %missing	E $\tilde{B}_w$	E $\tilde{B}_x$	$\tilde{V}_w$	$\tilde{V}_x$	E $\hat{B}_w$	E $\hat{B}_x$	$Var(\hat{B}_w)$	$Var(\hat{B}_x)$	$\frac{Var(\hat{B}_x)}{\tilde{V}_x}$
n=1000									
25% **Pr(M=1)=Φ(x-1)**	1.009	1.009	.008	.014	1.0098	1.0045	.008	.0109	.78
50% **Pr(M=1)=Φ(x)**	1.015	1.017	.013	.024	1.016	1.007	.013	.0131	.54
75% **Pr(M=1)=Φ(x+1)**	1.05	1.04	.038	.071	1.05	1.006	.0375	.021	.30

Table 2:
Monte Carlo Study: Comparison of the Efficient Estimator with Imputation approaches under MAR

Proportion of Missing Data	Estimator	B_w	B_x	V_w	V_x
n=1000					
25% **Pr(M=1)=Φ(x-1)**	Efficient Estimator Mean Imputation Multiple Imputation	1.0098 1.003 .95	1.0045 1.125 .97	.008 .008 .008	.0109 .0103 .0105
50% **Pr(M=1)=Φ(x)**	Efficient Estimator Mean Imputation Multiple Imputation	1.016 1.006 .88	1.007 1.27 .93	.013 .0138 .012	.0131 .009 .012
75% **Pr(M=1)=Φ(x+1)**	Efficient Estimator Mean Imputation Multiple Imputation	1.05 1.028 .797	1.006 1.377 .88	.0375 .0385 .023	.021 .008 .018

Table 3: Monte Carlo Study: Robustness of our estimator to departures for normality of W_i under MAR

	Complete Case Analysis				Adjusted Estimator				
% missing	E $\tilde{B}_w$	E $\tilde{B}_x$	$\tilde{V}_w$	$\tilde{V}_x$	E $\hat{B}_w$	E $\hat{B}_x$	$Var(\hat{B}_w)$	$Var(\hat{B}_x)$	$\frac{Var(\hat{B}_x)}{\tilde{V}_x}$
					n=1000				
25% missing									
Logistic	1.011	1.009	.008	.014	1.011	1.005	.008	.0111	.78
Uniform	1.01	1.01	.0075	.0135	1.01	1.002	.0075	.0106	.79
Binary	.999	1.01	.0148	.0084	.999	1.012	.0148	.0065	.77
Mult Imp.					.988	1.008	.0148	.0065	.77
50% missing									
Logistic	1.017	1.025	.0139	.025	1.017	1.016	.0138	.0134	.53
Uniform	1.017	1.017	.0124	.0238	1.018	.993	.0124	.0127	.54
Binary	1.005	1.012	.023	.013	.9995	1.03	.0028	.0068	.50
Mult Imp.					.982	1.01	.0023	.007	
75% missing									
Logistic	1.046	1.049	.038	.073	1.046	1.026	.038	.0216	.30
Uniform	1.039	1.04	.036	.07	1.04	.994	.036	.0204	.29
Binary	1.005	1.024	.056	.032	.987	1.07	.055	.0077	.24
Mult Imp.					.98	1.012	.056	.0103	

Table 4: Monte Carlo Study: Robustness of our amended estimator to misspecification of W_i function under MAR.

	Complete Case Analysis				Adjusted Estimator				
% missing	E $\tilde{B}_w$	E $\tilde{B}_x$	$\tilde{V}_w$	$\tilde{V}_x$	E $\hat{B}_w$	E $\hat{B}_x$	$Var(\hat{B}_w)$	$Var(\hat{B}_x)$	$\frac{Var(\hat{B}_x)}{\tilde{V}_x}$
					n=1000				
					(Omitted G; Correlation(G,X)=.5)				
50% missing	1.02	1.02	.0126	.0327	1.023	1.005	.0125	.0192	.59
					(Omitted X^2: Complete Case versus Original Efficient Estimator)				
50% missing	1.013	1.014	.077	.0144	1.012	1.26	.0076	.0077	
					(Omitted X^2: Complete Case versus Adjusted Efficient Estimator)				
50% missing	1.013	1.014	.077	.0144	1.013	1.014	.077	.0104	.70

Table 5
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 6
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