Laitila, Thomas; Wang, Lisha

Working Paper

A Two-Step Estimator for Missing Values in Probit Model Covariates

Working Paper, No. 3/2015

Provided in Cooperation with:
Örebro University School of Business

Suggested Citation: Laitila, Thomas; Wang, Lisha (2015) : A Two-Step Estimator for Missing Values in Probit Model Covariates, Working Paper, No. 3/2015, Örebro University School of Business, Örebro

This Version is available at:
https://hdl.handle.net/10419/244507

WORKING PAPER

3/2015

A Two-Step Estimator for Missing Values in Probit Model Covariates

Lisha Wang and Thomas Laitila
Statistics

ISSN 1403-0586

http://www.oru.se/Institutioner/Handelshogskolan-vid-Orebro-universitet/Forskning/Publikationer/Working-papers/
Örebro University School of Business
701 82 Örebro
SWEDEN

A Two-Step Estimator for Missing Values in Probit Model Covariates

LISHA WANG
Department of Statistics, Örebro university, SE-70182 Örebro
lisha.wang@oru.se

THOMAS LAITILA
Research and Development Deparment, Statistics Sweden, SE-70189 Örebro
thomas.laitila@oru.se

April 27, 2015

Abstract

This paper includes a simulation study on the bias and MSE properties of a two-step probit model estimator for handling missing values in covariates by conditional imputation. In one smaller simulation it is compared with an asymptotically efficient estimator and in one larger it is compared with the probit ML on complete cases after listwise deletion. Simulation results obtained favors the use of the two-step probit estimator and motivates further developments of the methodology.

Keywords binary variable, imputation, OLS, heteroskedasticity

1 Introduction

A major issue in applied research is the effect of covariates on study variables. Often the topic considered involves a binary, two-valued dependent variable. Examples are if a new method of teaching improves performance in later courses or not (Greene (2008)), and if a married woman participates in the labor force or not (Chib & Greenberg (1998)). One of the two most frequently employed statistical models for analysis of such study variables is the probit model (Bliss (1934)). This model is nonlinear in parameters and is usually estimated with maximum likelihood (ML), i.e. using the probit ML estimator (e.g. Hayashi (2000)).

Frequent in applied work, one or several model covariates are associated with missing values. Deletion of all observations having missing values, so called list wise deletion, is not to recommend, but is often done for practical reasons. Little & Rubin (1987) notes that using only Complete-Case (CC) observations causes loss of information unless they are MCAR. Enders (2010) also find that deletion of observations produces distorted estimates in most situations. List wise deletion is thus not a statistically efficient way of using available data, and alternative methods have been suggested. Little & Rubin (1987) discuss in detail on combination of different types of missing data analyses, and expanded coverage of bayesian methodology. Schafer (1997) discussed statistical properties of different kinds of methods dealing with incomplete multivariate data with continuous variable, categorical variable or both.

Available methods for handling missing values in covariates depend on the form of model to estimate. ML estimation may be used if the model is fully parametric and involves models for the covariates as well (e.g. Gourieroux & Monfort (1981)). For the probit model, Conniffe & O′Neill (2011) propose an asymptotically efficient estimator under missing values of covariate variables. Their estimator correspond to a first round iteration of the method of scoring algorithm for the ML estimator using consistent initial estimates.

Our objective in this paper is to study a related two-step estimation procedure by means of simulation. The two-step estimator of the probit model corresponds to an adaption of the generalized least squares estimator suggested by Dagenais (1973) for linear regression with missing covariate values. It can also be interpreted as a feasible limited information ML estimator since a distributional assumption is utilized in a second probit ML estimation step.

The main idea of the estimation procedure is explained in terms of estimation of a linear regression model in the first part of the next section. In the second part, the idea is applied to the probit model and consistency properties are addressed. Section 2 also includes a numerical comparison of the properties of the estimator with those of the Conniffe & O′Neill (2011) estimator. In Section 3 the two-step estimation procedure is compared with probit ML after list wise deletion. The comparison is made by simulation using a real data set for generation of populations. A summary of results and ideas for future research are given in the final section.

2 Probit ML estimation with missing data

2.1 Imputation in the linear model

Consider the linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \tag{1}$$

where y_i is the dependent variable, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})'$ is the vector of regressors, $\boldsymbol{\beta}$ is the associated vector of regression coefficients, and $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$ is a random disturbance term independently distributed with zero mean and variance σ_i^2 conditionally on the regressors, i.e. $\varepsilon_i|\mathbf{x}_i \sim (0, \sigma_i^2)$.

Suppose observations of the regressors includes some missing values. Introduce a $K \times K$ diagonal matrix R_i whose k:th element equals 1 if x_{ik} is observed and zero if the value is missing. Also, let $\bar{R}_i = I - R_i$, where I is the identity matrix.

Let $\mathbf{z}_i$ be an additional vector of variables used for modeling regressor variables. Write $E(\mathbf{x}_i|R_i\mathbf{x}_i, \mathbf{z}_i) = R_i\mathbf{x}_i + E(\bar{R}_i\mathbf{x}_i|R_i\mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}_{ri} + \mu_{\bar{r}i}$. Note that the mean of variables with missing values are made conditional on observed variables. Thus, if the set of observed variables is different, the conditional expectation of a missing variable may be different. When modeling for missing values of a variable in different observations, the models may therefore be different. Also note that the vector $\mu_{\bar{r}i}$ contains zeros for variables with observed values. Missing values are in $\mathbf{x}_{ri}$ represented by zeros.

Using the conditional mean values in place of missing values yields the imputed regressor vector

$$\mathbf{x}_{\bullet i} = \mathbf{x}_{ri} + \mu_{\bar{r}i}$$

which gives the imputation errors $M_i = \mathbf{x}_i - \mathrm{x}_{\bullet i} = \bar{R}_i(\mathbf{x}_i - \mu_{\bar{r}i})$. These errors have conditional mean $E(M_i|\mathbf{x}_{ri}, \mathbf{z}_i) = 0$ and covariance $Cov(M_i|\mathbf{x}_{ri}, \mathbf{z}_i) =$

$\bar{R}_i Cov(\mathbf{x}_i - \mu_{\bar{r}i}|\mathbf{x}_{ri}, \mathbf{z}_i)\bar{R}_i = \Sigma_{\bar{r}i}$. In this covariance matrix variances and covariances corresponding to observed variables are zero. It is also here assumed that imputation errors between different units are independent with e.g. $E(M_i M_j^t|x_{ri}, x_{rj}, z_i, z_j) = 0$ $(i \neq j)$.

For the linear regression considered, the imputations yield the model

$$y_i = \mathbf{x}'_{\bullet i}\boldsymbol{\beta} + \eta_i \tag{2}$$

where the disturbance term $\eta_i = \varepsilon_i + M_i'\boldsymbol{\beta}$ has mean $E(\eta_i|\mathbf{x}_{ri}, \mathbf{z}_i) = 0$ and variance $\omega_i = V(\eta_i|\mathbf{x}_{ri}, \mathbf{z}_i) = \beta'\Sigma_{\bar{r}i}\beta + \sigma^2$.

For the moment, assuming knowledge of $\mu_{\bar{r}i}$ and $\Sigma_{\bar{r}i}$ there are at least two alternatives for estimation of model (2) (Dagenais (1973), Gourieroux & Monfort (1981)). The first is to apply OLS and correct for heteroskedasticity in variance estimation, where the OLS estimator has conditional covariance matrix

$$(\sum_{i=1}^{N} \mathbf{x}_{\bullet i}\mathbf{x}'_{\bullet i})^{-1} \sum_{i=1}^{N} \omega_i \mathbf{x}_{\bullet i}\mathbf{x}'_{\bullet i} (\sum_{i=1}^{N} \mathbf{x}_{\bullet i}\mathbf{x}'_{\bullet i})^{-1}$$

A second alternative is to make use of the OLS estimate and derive variance estimates $\hat{\omega}_i = \hat{\beta}'_{OLS}\hat{\Sigma}_{\bar{r}i}\hat{\beta}_{OLS} + \hat{\sigma}^2$ and apply feasible WLS

$$\hat{\beta}_{FWLS} = (\sum_{i=1}^{N} (1/\hat{\omega}_i)\mathbf{x}'_{\bullet i}\mathbf{x}_{\bullet i})^{-1} \sum_{i=1}^{N} (1/\hat{\omega}_i)\mathbf{x}'_{\bullet i} y_i$$

As an alternative to running OLS on model (2) for deriving an initial estimate of β, OLS can be applied to the subset of complete cases.

In practice the conditional means $\mu_{\bar{r}i}$ and covariance matrices $\Sigma_{\bar{r}i}$ are unknown and have to be estimated. Following (Conniffe & O′Neill, 2011) separate models for the covariates have to be specified and estimated on available data.

2.2 Imputation in the probit model

The above approach is readily applicable to missing data problems in estimation of the binary probit model. Replace the dependent variable in model (1) with the latent variable y_i^* and set $\sigma^2 = 1$. Observations are only made on an indicator variable $y_i = 1(y_i^* > 0)$, stating whether the latent variable is positive or not. Assuming the error $\eta_i = \varepsilon_i + M_i'\boldsymbol{\beta}$ to be normally distributed, then

$$P(y_i = 1|x_{\bullet i}) = \Phi(\frac{\mathbf{x}'_{\bullet i}\beta}{\sqrt{\beta'\Sigma_{\bar{r}i}\beta + 1}}) \tag{3}$$

The model given in (3) is related to Burr (1988) who studies Berkson regressor measurement errors in probit modelling of dose/response experiments. For identifiability she finds it necessary to either know the regressor slope coefficient or the measurement error variance, which neither are usually known.

Here the model (3) is in the context of imputation, and the imputation error covariance matrix is assumed at least estimable. Thus, given knowledge of the conditional means $\mu_{\bar{r}i}$ and the covariances $\Sigma_{\bar{r}i}$ this model can be estimated with ML.

Conniffe & O′Neill (2011) suggest an efficient estimator by considering simultaneous estimation of the probit model (3) and the parameters in the models for

the missing variables, assuming a simultaneous distribution for the response and explanatory variables. Their idea is to estimate the parameters with ML using complete cases in a first step. These are used as initial, consistent estimates in a second step. For the second step, they set up the log-likelihood for all observations, including those with missing values. Their efficient estimator is then defined as the outcome of the first round of the methods of scoring algorithm for the MLE.

A two-step probit estimator is studied in the present paper. In the first step, the probit ML estimator is used on the complete cases only (CC probit), yielding estimates $\hat{\beta}_{CC}$. Then estimates of the conditional means and covariances, i.e. $\hat{\mu}_{\bar{r}i}$ and $\hat{\Sigma}_{\bar{r}i}$, respectively, are obtained from estimates of appropriate models. In the second step, the imputed regressor vectors are weighted into $\hat{\mathbf{x}}_{\bullet i}(\hat{w}) = \hat{\mathbf{x}}_{\bullet i}/\sqrt{\hat{w}_i} = (\mathbf{x}_{ri} + \hat{\mu}_{\bar{r}i})/\sqrt{\hat{\beta}'_{CC}\hat{\Sigma}_{\bar{r}i}\hat{\beta}_{CC} + 1}$ and probit ML estimation is applied for estimation of the model

$$P(y_i = 1|\hat{\mathbf{x}}_{\bullet i}(\hat{w})) = \Phi(\hat{\mathbf{x}}_{\bullet i}(\hat{w})'\beta) \tag{4}$$

yielding the two-step probit estimator $\hat{\beta}_{2S}$.

Two-step estimators can be shown consistent under appropriate conditions (e.g. Wooldridge (2002), Sec. 13.10). Consider imputations made and estimated variances ($\hat{w}_i$) as functions of an estimate of a parameter vector θ, i.e. $\hat{\mathbf{x}}_{\bullet i}(\hat{w}) = \mathbf{x}_{\bullet i}(w : \hat{\theta})$, such that $\mathbf{x}_{\bullet i}(w) = \mathbf{x}_{\bullet i}(w : \theta)$. Let $\hat{l}(b) = (1/n)log(\hat{L}(b))$ denote the log likelihood based on (4), and let $l(b) = (1/n)\sum y_i log\Phi(\mathbf{x}_{\bullet i}(w)'b) + (1-y_i)log(1-\Phi(\mathbf{x}_{\bullet i}(w)'b))$. By Taylor expansion $\hat{l}(b) - l(b) = S_n(b, \theta^*)(\hat{\theta} - \theta)$, where $S_n(b, \theta^*)$ is a vector of first order derivatives wrt θ evaluated at a point between $\hat{\theta}$ and θ. If $S_n(b, \theta^*)$ is bounded a.s. and $(\hat{\theta} - \theta) \to 0$ a.s., then $\hat{l}(b) - l(b) \to 0$ a.s. Lemma 3 in Amemiya (1973) can then be used to obtain consistency of the two-step estimator since $l(b)$ is the log-likelihood for the standard probit model.

2.3 Comparison with an efficient estimator

Table 1 reports simulation results on bias and standard error for the two-step probit estimator. The setup of the simulation study equals the one reported in Conniffe & O′Neill (2011), and the results in their table are reproduced. The purpose is to compare the two-step probit estimator with their efficient estimator.

The model is

$$y_i^* = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \tag{5}$$

where $x_{1i} \sim N(0, 1)$, $x_{2i} = cx_{1i} + u_i$, $u_i \sim N(0, \sigma)$, and $\varepsilon_i \sim N(0, 1)$. The parameters in the model are fixed at $(\beta_1, \beta_2, c, \sigma)$=(1,1,1,1). Missing values on the second regressor, x_{i2}, are generated using the model $Pr(R_{2i} = 1 \mid x_{i1}) = \Phi(x_{i1} - b)$, where b=-1, 0, or 1.

First, the results obtained here for the CC probit estimator are similar to those reported by Conniffe & O′Neill (2011), suggesting comparability over the two studies. There is a tendency to smaller variance estimates for β_1. The results for the two-step probit estimator compare well with those of the Conniffe & O′Neill (2011) estimator (EE), especially for the lower proportions of missing values.

Table 2 presents the means of the probit ML variance estimates in the second step. They show the probit ML variance estimator of the variance of β_1 give too large estimates, while the one for the variance of β_2 gives too low estimates.

	β_1	β_2
	25%, $Pr(M=1)=\Phi(x-1)$	
two-step probit	1.0076	1.0028
	(.0103)	(.009)
EE (Conniffe & O'Neill′s)	1.0045	1.0098
	(.0109)	(.008)
CC probit	1.008	1.003
	(.012)	(.008)
CC (Conniffe & O'Neill′s)	1.009	1.009
	(.014)	(.008)
	50%, $Pr(M=1)=\Phi(x)$	
two-step probit	1.009	1.007
	(.0123)	(.015)
EE (Conniffe & O'Neill′s)	1.007	1.016
	(.0131)	(.013)
CC probit	1.008	1.009
	(.019)	(.013)
CC (Conniffe & O'Neill′s)	1.017	1.015
	(.024)	(.013)
	75%, $Pr(M=1)=\Phi(x+1)$	
two-step probit	1.042	0.993
	(.021)	(.042)
EE (Conniffe & O'Neill′s)	1.006	1.05
	(.021)	(.0375)
CC probit	1.029	1.025
	(.041)	(.036)
CC (Conniffe & O'Neill′s)	1.04	1.05
	(.071)	(.038)

Table 1: Estimated means and variances (in parenthesis) of estimators of model (5) for different levels of missing values of variable w (1000 replications, n=1000)

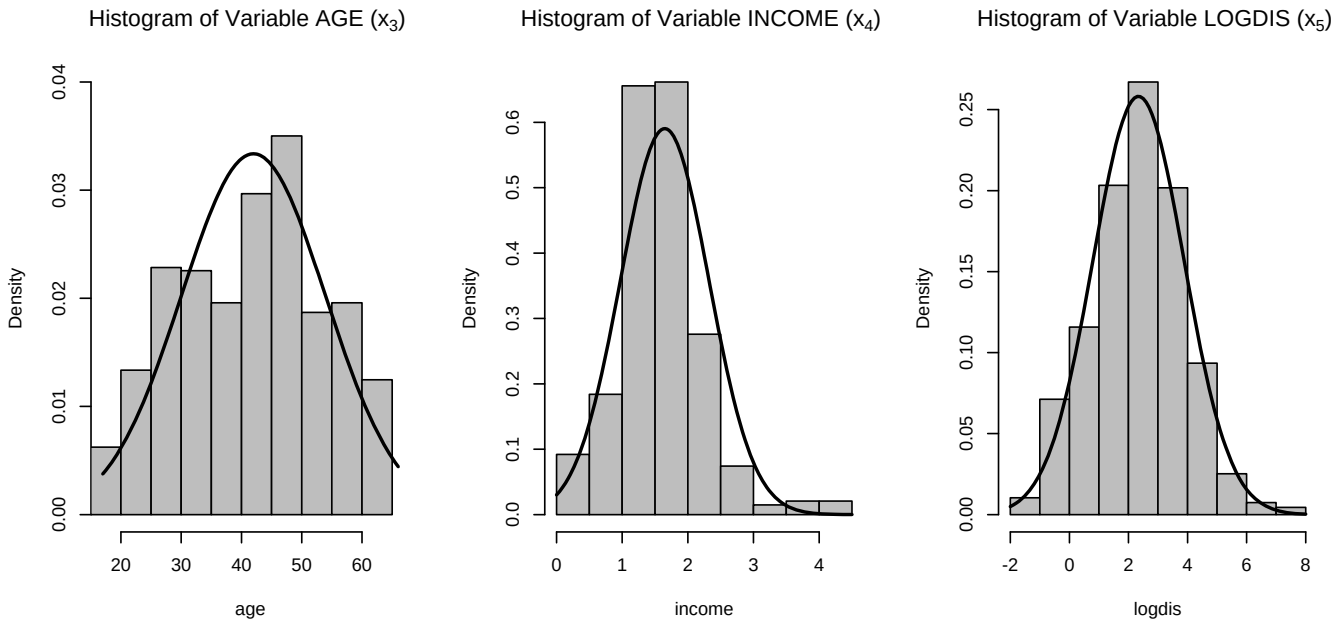
	β_1	β_2
	25%, $Pr(M=1)=\Phi(x-1)$	
two-step probit	0.0111 (.076)	0.0072 (-.2004)
CC probit	0.0133 (.083)	0.0079 (-.032)
	50%, $Pr(M=1)=\Phi(x)$	
two-step probit	0.0138 (.122)	0.0105 (-.285)
CC probit	0.0200 (.073)	0.0129 (.030)
	75%, $Pr(M=1)=\Phi(x+1)$	
two-step probit	0.0236 (.138)	0.0223 (-.527)
CC probit	0.0413 (.011)	0.0356 (.031)

Table 2: Means and relative bias (in parenthesis) of probit ML variance estimates for the two-step probit estimator and the CC probit estimator in Table 1

Variable Name	Value	Frequency	Proportion
(y) DL	1=holding a driving licence	612	90.8%
	0=no driving licence	62	9.2%
(x_1) gender	1=female	304	45.1%
	0=male	370	54.9%
(x_2) Cs	1=married or sambo	418	62.0%
	0=single	256	38.0%

Table 3: Summary of Categorical Variables

Figure 1: Histogram of Continuous Variables



3 Simulation Study

3.1 Data and models

To illustrate the properties of the two-step probit estimator, a simulation experiment is conducted based on a real dataset with 674 observations. This dataset contains a binary variable describing whether or not a person holds a driving license. The variable is here denoted DL. Other variables in the data are age, distance between work and home, yearly income, civil status (denoted as Cs hereafter) and a female gender dummy variable.

Table (3) shows descriptives of the categorical variables DL (y), gender (x_1) and Cs (x_2). Figure (1) shows histograms of age (x_3), income (x_4) and the logarithm of distance between home and work (logdis, x_5). Derived variables used are age squared $(x_6 = x_3^2)$ and age times gender $(x_7 = x_1 * x_3)$.

In the simulations, DL is treated as the dependent variable, while Cs (x_2), age (x_3), income (x_4) and logdis (x_5) are used as regressors. This model was fitted to the data set and used as the true model in the simulations. Additional variables were tested using a backward selection procedure (Cox & Snell, 1981), and the likelihood ratio test was used for selection between models. The ML estimates of the final model are displayed in Table (4).

	Intercept	Cs (x_2)	age (x_3)	income (x_4)	logdis (x_5)
estimates	-0.3517	0.2770	0.0120	0.6035	0.0772
Standard error	0.3170	0.1576	0.0067	0.1450	0.0476

Table 4: ML estimates of probit model regressing y on x_2, x_3, x_4 and x_5, $R_p^2 = 0.227$ [1]

	Intercept	gender (x_1)	agegender (x_7)
estimates	1.2513	1.2656	-0.0792
Standard error	0.0875	0.3534	0.0089

Table 5: ML estimates of probit model regressing Cs (x_2) on x_1 and x_7, $R_p^2 = 0.59$

In practice, modelling of variables with missing values can be done using the data set at hand as well as additional variables available. Careful modeling requires a controlled process where different formulations of the model are evaluated and tested against each other. Such procedures are not possible to implement in this simulation study.

The models used for imputation in the simulation are obtained from analyses of the data set available. Two variables, Cs and Income, are considered for missing values in the simulation. The imputations for Cs are made using the probit model reported in Table (5). The linear regression model reported in Table (6) are used for imputations of Income.

In the simulations, before imputing missing values in an iteration, the models shown in tables (5) and (6) are re-estimated on the complete cases in that iteration.

3.2 Simulation setup

The general technique used in the simulation, is to first draw a sample from the population (the data set of 674 observations), and then generate the dependent variable using the fitted model in Table (4). Then missing values are generated at random followed by imputing values using the models in tables (5) and (6) fitted to the complete cases. Thereafter the two-step probit estimator is applied. In more detail, the simulations are made in the following steps.

1. A simple random sample with size n is selected from the population. For each selected unit, the binary response variable y_i is generated from a random draw from the Bernoulli ($\Phi(X'\beta)$) distribution. Here X is the vector of variables and β is the parameter estimates in Table (4). The intercept is adjusted to yield an approximately 50/50 ratio of 0's and 1's.

2. Missing values in explanatory variables (Cs or income, or both) are generated by taking a simple random sample. Two independent simple random samples are taken to generate cases with missing values in both variables.

3. For CC estimation, probit regression is applied on the subset of observation with non-missing values. This estimator is denoted $\tilde{\beta}_{CC}$. The

[1] The pseudo-goodness-of-fit (R_p^2) utilized here is proposed by Laitila (1993). For probit case, $R_p^2 = \hat{\beta}'_{ML}\hat{\Sigma}_{\mathbf{x}}\hat{\beta}_{ML}/1 + \hat{\beta}'_{ML}\hat{\Sigma}_{\mathbf{x}}\hat{\beta}_{ML}$, where $\hat{\beta}_{ML}$ is the MLE of the probit model and $\hat{\Sigma}_{\mathbf{x}}$ is the covariance matrix of the regressors.

	Intercept	age (x_3)	logdis (x_5)	age^2 (x_6)	agegender (x_7)
estimates	-0.8577	0.1074	0.0602	-0.000984	-0.0144
Standard error	0.2306	0.0116	0.0131	0.00014	0.00093

Table 6: OLS estimates of linear model regressing income (x_4) on x_3, x_5, x_6 and x_7, where $R^2 = 0.41$

models used for imputations of Cs and income, respectively, are estimated on complete cases.

4. Missing values are imputed. For the Cs variable, the imputation error variance is estimated by $\Lambda_i(1-\Lambda_i)$ where Λ_i is the estimated probability of $\text{Cs}_i = 1$. For income, the error variance is estimated as the residual variance in the fit of the model (6) to complete cases. The covariance between imputation errors is set to zero.

5. The regressor vector $\mathbf{x}_{\bullet i}$ is divided with $\sqrt{\tilde{\beta}'_{CC}\tilde{\Sigma}_{\bar{r}i}\tilde{\beta}_{CC} + 1}$ or with 1, depending on if it contains imputed elements or not. Resulting regressor vector is denoted $\tilde{\mathbf{x}}_i$.

6. The two-step probit estimate is obtained by probit ML regressing the generated y_i on $\tilde{\mathbf{x}}_i$.

7. Steps 1 - 6 are repeated 1000 times.

3.3 Results

Figure 2 illustrates the comparison of the absolute values of the bias estimates of the CC probit and the two-step probit estimators. The case considered is missing values in the binary civil status variable Cs. The sample size varies from 100 to 500 and the response rate varies from 30% to 90%.

The bias estimates for the two-step procedure keep staying on a lower platform than that of the CC probit estimator for all variables except for Cs. For larger sample sizes and higher response rates, the bias estimates for the two estimators are close to each other. With exception for the Cs variable, the figure shows on more pronounced smaller biases for the two-step probit estimator when the response rate is small and, the difference increases with a decreasing sample size.

For the Cs variable this pattern is not observed for the lower response rate. The bias for the two-step probit estimator is either smaller than the one for the CC probit estimator, or the biases for the two estimators are close.

Figure 3 illustrates corresponding comparison of MSE estimates for the case considered in Figure 2. The MSEs of the estimators for all four parameter estimates have the same pattern. The MSE estimates for the two-step probit estimator are smaller than those of the CC probit estimator for all sample sizes and response rates. The MSEs obtained for the two estimators are close when both response rate and sample size are high. The MSE of the CC probit estimator increases more rapidly than that of the two-step probit estimator with the decrease of the response rate and the sample size. Results for the estimators of the intercept is not shown, but plotted bias and MSE estimates yield surfaces similar to the ones for the income variable.

Additional bias and MSE results are shown in tables 7 and 8 for the $n = 100$ case. The results in the upper third are for missing values in Cs, the middle

Figure 2: Bias comparison in CC and two-step procedure
when missing value occurs in Cs

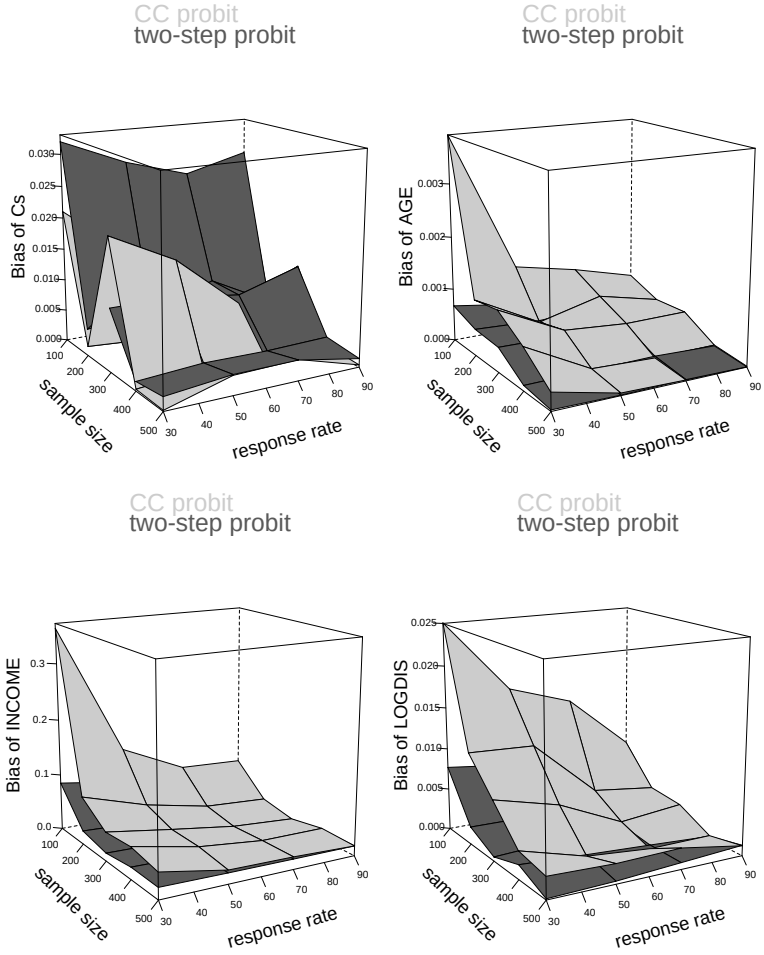
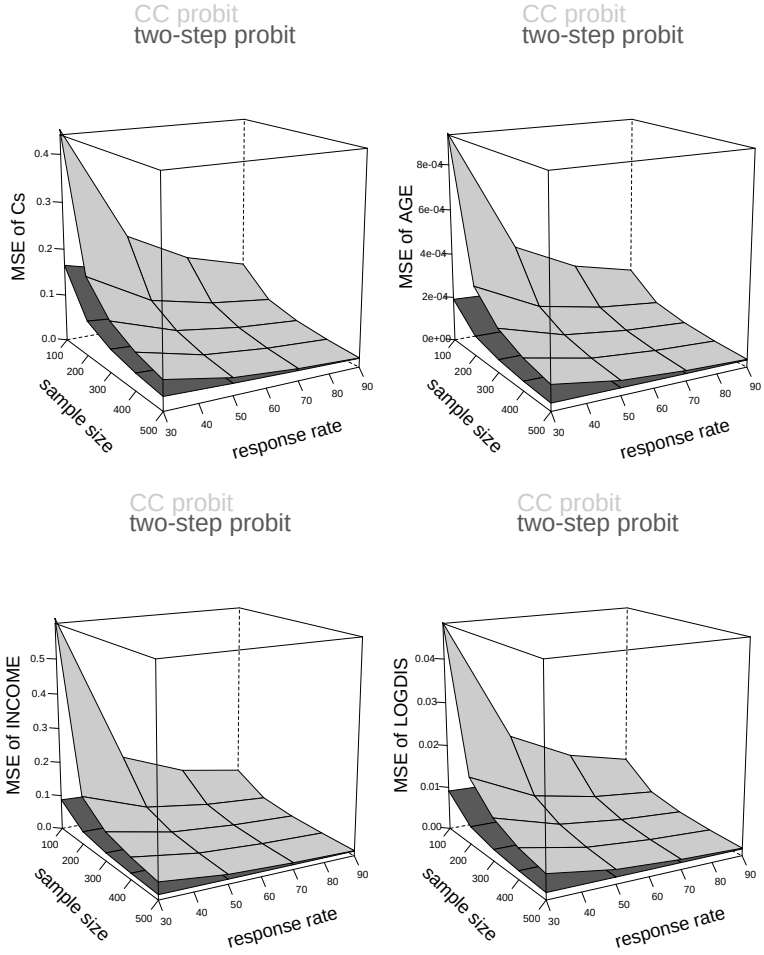


Figure 3: MSE comparison in CC and two-step procedure
when missing value occurs in Cs



third for missing values in income, and the lower third for missing values in both variables. The bias estimates generally increase with the increase of proportion of missing values. The value (or absolute value) of bias estimates for the two-step probit estimator are smaller than those for the CC probit estimator. This is observed in all cases except for the Cs variable when missing values occur only in that variable. Correspondingly, the MSE estimates gets larger when the proportion of missing values gets larger. Here, the MSE estimates for the two-step probit estimator are smaller than those for the CC probit estimator in all cases without exception.

In Table 9, coverage rates are shown for 95% confidence intervals of $\hat{\beta}_{CC}$ and $\hat{\beta}_{2S}$. The coverage rates are calculated as $\sum_{i=1}^{1000} I_i/1000$ where the indicator $I_i = 1(\hat{\beta}_i - z_{\alpha/2}SE_{\hat{\beta}_i} < \beta < \hat{\beta}_i + z_{\alpha/2}SE_{\hat{\beta}_i})$, stating whether the 95% confidence interal covers β or not. The standard error used is the one obtained from the probit ML estimator in the second step. Results are for the sample size $n = 100$ and show that the two-step probit estimator works comparably well with the CC probit estimator. Using separate t-tests shows that 68.9% of the coverage rates of the two-step procedure do not significantly differ from 95%.

To explore how the two-step procedure works in a less balanced case, the percentage of 1's of the response variable y is decreased from 0.5 to 0.3. The comparison of bias and MSE of the estimators of the CC probit model and the two-step procedure is displayed in Table 10. Another set of simulations is made where the latent linear regression model behind the binary response have a 50% increase in R^2, from 0.227 to 0.341. Results are presented in Table 11. The results obtained in both additional simulations are similar to those in tables 7 and 8, where the MSEs of the two-step probit estimator for all the parameters are smaller than that of the CC probit estimator, and the corresponding biases of the two-step probit estimator are smaller than that of the CC probit estimator except for a few cases.

4 Discussion

Listwise deletion of observation with missing values implies loss of precision in estimation. This paper consider estimation of a probit model and studies the potentials of a two-step probit estimation procedure, where missing values are imputed with estimated conditional expectations and weighted for additional imputation error variance.

Simulation results suggest the two-step probit estimator to reduce bias and increase efficiency compared with probit ML using complete cases only. The simulation study shows that the two-step probit estimator produces more accurate estimates in terms of smaller bias and MSE. Smaller MSE estimates are obtained for all sample sizes and response rates considered. Only when imputations are made for a missing dichotomous regressor, biases are observed larger than probit on complete cases. The difference in bias is small, however.

The two-step procedure keeps a lower MSE even at the larger sample sizes ($n = 500$) and the larger response rates considered. Although the general results are in favor of the two-step probit estimator, the greatest gains are observed for small sample sizes and high rates of missing values.

The two-step probit estimator provide a model in the second step which deviates from the normality assumption behind the probit model, unless regressor variables are normally distributed. This leads to a model misspecification leading to potential inconsistency of the two-step probit ML estimator. The effects of this misspecification were not observed as severe as expected, and

the increase in bias is well compensated with reduction in variance, yielding a lower MSE.

This robustness property of the two-step probit estimator were also observed by Conniffe & O′Neill (2011). One of their simulations are here replicated for the two-step probit estimator and results shows it to compare well with the Conniffe & O′Neill (2011) estimator.

One advantage of the two-step probit estimator is its simplicity in application, standard software can be used. Our purpose here is to study the potentials in using the estimator with respect to bias and MSE. Results are in summary most promising and further study of properties of the estimator is motivated. There are at least two topics to consider. One is variance estimation. Using the standard errors from the second probit step gave almost correct coverage rates of 95% parameter confidence intervals in the $n = 100$) case, although standard errors are inconsistent (Murphy & Topel (1985)). A potential improvement is bootstrapping. The extra uncertainty come from the initial parameter estimates and the estimates of the conditional means. Bootstrapping these estimates, gives a bootstrap sample of the two-step probit estimator which can be utilized for measuring the extra variability.

A second topic for further study are the asymptotic properties of the two-step probit estimator. Of special interest here is consistency, or rather, the size of the asymptotic bias if assumptions of normally distributed regressors are not met. When the two-step probit estimator is preceeded by an initial probit estimate on complete cases it may be possbile to use a Hausman type of model misspecification test. Such a test could guard against cases when the normality assumption behind the probit model in the second step is grossly violated.

Response Rate			Variable				
Cs	income		Intercept	Cs (x_2)	age (x_3)	income (x_4)	logdis (x_5)
70%	100%	CC	-0.1847	-0.0199	0.0010	0.0774	0.0139
		two-step	-0.1282	-0.0248	0.0006	0.0643	0.0075
50%	100%	CC	-0.2910	-0.0130	0.0012	0.1296	0.0164
		two-step	-0.1270	-0.0278	0.0005	0.0680	0.0068
30%	100%	CC	-0.7991	-0.0210	0.0039	0.3617	0.0251
		two-step	-0.1610	-0.0319	0.0007	0.0850	0.0077
100%	70%	CC	-0.1954	0.0212	0.0006	0.0841	0.0075
		two-step	-0.0994	0.0017	-0.0001	0.0560	0.0031
100%	50%	CC	-0.3070	0.0275	0.0013	0.1208	0.0143
		two-step	-0.0758	0.0080	0.0000	0.0317	0.0046
100%	30%	CC	-0.7414	0.0143	0.0047	0.2810	0.0357
		two-step	-0.0376	0.0254	0.0005	-0.0171	0.0070
70%	70%	CC	-0.2884	-0.0029	0.0008	0.1454	0.0094
		two-step	-0.0818	-0.0050	-0.0003	0.0622	0.0004
60%	60%	CC	-0.5060	0.0276	0.0026	0.2047	0.0185
		two-step	-0.0916	-0.0034	-0.0003	0.0655	0.0004
50%	50%	CC	-0.9352	-0.0430	0.0035	0.4456	0.0406
		two-step	-0.1216	-0.0158	-0.0005	0.0920	0.0004

Table 7: Bias of estimates in CC probit and two-step Procedure when n=100

Response Rate			Variable				
Cs	income		Intercept	Cs (x_2)	age (x_3)	income (x_4)	logdis (x_5)
70%	100%	CC	0.7058	0.1412	0.0003	0.1161	0.0131
		two-step	0.4750	0.1146	0.0002	0.0784	0.0087
50%	100%	CC	1.1068	0.2089	0.0004	0.1864	0.0200
		two-step	0.4860	0.1329	0.0002	0.0807	0.0088
30%	100%	CC	3.5122	0.4507	0.0009	0.6141	0.0482
		two-step	0.5344	0.1649	0.0002	0.0872	0.0091
100%	70%	CC	0.7111	0.1429	0.0003	0.1184	0.0130
		two-step	0.4489	0.0983	0.0002	0.0954	0.0087
100%	50%	CC	1.1241	0.2137	0.0004	0.1868	0.0197
		two-step	0.4435	0.1016	0.0002	0.1121	0.0089
100%	30%	CC	2.9505	0.4603	0.0009	0.4851	0.0442
		two-step	0.4514	0.1079	0.0002	0.1437	0.0093
70%	70%	CC	1.1543	0.2196	0.0004	0.2023	0.0205
		two-step	0.4596	0.1174	0.0002	0.0921	0.0087
60%	60%	CC	1.9603	0.3347	0.0006	0.3297	0.0318
		two-step	0.4706	0.1275	0.0002	0.0968	0.0088
50%	50%	CC	4.7627	0.7132	0.0013	0.8766	0.0687
		two-step	0.5018	0.1421	0.0002	0.1070	0.0091

Table 8: MSE of estimates in CC probit and two-step Procedure when n=100

Response Rate			Variable				
Cs	income		Intercept	Cs (x_2)	age (x_3)	income (x_4)	logdis (x_5)
70%	100%	CC	95.4%	94.1%	94.6%	94.7%	95.0%
		two-step	95.1%	94.9%	93.0%	93.6%	94.6%
50%	100%	CC	95.2%	94.8%	93.7%	95.5%	93.9%
		two-step	94.2%	93.7%	93.0%	93.6%	94.2%
30%	100%	CC	97.2%	93.9%	94.6%	95.6%	94.6%
		two-step	94.4%	93.9%	93.3%	94.5%	93.9%
100%	70%	CC	94.6%	93.2%	93.7%	95.1%	95.3%
		two-step	95.1%	94.9%	93.2%	94.9%	95.2%
100%	50%	CC	95.0%	94.2%	94.1%	95.2%	93.3%
		two-step	93.9%	94.0%	92.8%	95.6%	94.4%
100%	30%	CC	95.5%	94.4%	94.8%	95.7%	94.5%
		two-step	93.5%	94.0%	93.1%	93.3%	92.7%
70%	70%	CC	94.2%	92.5%	94.0%	93.7%	93.6%
		two-step	95.0%	93.1%	95.3%	94.0%	94.2%
60%	60%	CC	95.1%	94.4%	95.8%	95.2%	94.9%
		two-step	94.8%	93.7%	95.7%	93.7%	94.3%
50%	50%	CC	96.4%	95.1%	96.0%	96.2%	95.2%
		two-step	94.2%	92.3%	94.7%	93.1%	94.0%

Table 9: Coverage Rate of 95% CI of estimates in CC probit and two-step Procedure when n=100

Response Rate			Variable				
Cs	income		Intercept	Cs (x_2)	age (x_3)	income (x_4)	logdis (x_5)
70%	100%	CC	-0.2390	-0.0059	-0.0003	0.1218	0.0023
			(0.8522)	(0.1630)	(0.0003)	(0.1311)	(0.0145)
		two-step	-0.1239	-0.0248	-0.0005	0.0791	0.0016
			(0.5406)	(0.1290)	(0.0002)	(0.0824)	(0.0096)
50%	100%	CC	-0.4166	-0.0049	0.0010	0.1758	0.0100
			(1.4306)	(0.2476)	(0.0005)	(0.2144)	(0.0226)
		two-step	-0.1348	-0.0261	-0.0005	0.0847	0.0022
			(0.5573)	(0.1499)	(0.0002)	(0.0856)	(0.0097)
30%	100%	CC	-0.8632	-0.0481	0.0032	0.3236	0.0423
			(3.8548)	(0.5494)	(0.0011)	(0.5521)	(0.0541)
		two-step	-0.1905	-0.0245	-0.0002	0.1047	0.0030
			(0.6176)	(0.1867)	(0.0002)	(0.0947)	(0.0101)
100%	70%	CC	-0.2436	0.0207	0.0001	0.0943	0.0115
			(0.8474)	(0.1643)	(0.0003)	(0.1229)	(0.0146)
		two-step	-0.1106	-0.0148	-0.0006	0.0647	0.0076
			(0.5211)	(0.1109)	(0.0002)	(0.1000)	(0.0096)
100%	50%	CC	-0.3739	0.0000	0.0001	0.1580	0.0188
			(1.3730)	(0.2496)	(0.0004)	(0.2091)	(0.0223)
		two-step	-0.0823	-0.0142	-0.0007	0.0518	0.0075
			(0.5159)	(0.1150)	(0.0002)	(0.1215)	(0.0098)
100%	30%	CC	-0.9408	-0.0579	0.0034	0.3491	0.0471
			(4.2221)	(0.5829)	(0.0011)	(0.6033)	(0.0551)
		two-step	-0.0675	0.0033	-0.0002	0.0183	0.0101
			(0.5324)	(0.1217)	(0.0002)	(0.1582)	(0.0103)
70%	70%	CC	-0.3635	-0.0089	0.0005	0.1526	0.0132
			(1.3933)	(0.2547)	(0.0005)	(0.2128)	(0.0230)
		two-step	-0.1266	-0.0386	0.0004	0.0652	0.0036
			(0.5383)	(0.1333)	(0.0002)	(0.0957)	(0.0096)
60%	60%	CC	-0.6518	-0.0325	0.0017	0.2617	0.0301
			(2.5460)	(0.3935)	(0.0007)	(0.3901)	(0.0381)
		two-step	-0.1482	-0.0361	0.0004	0.0732	0.0035
			(0.5564)	(0.1443)	(0.0002)	(0.1022)	(0.0098)
50%	50%	CC	-1.3937	-0.1728	0.0046	0.5388	0.0798
			(9.5247)	(1.1349)	(0.0022)	(2.3155)	(0.1345)
		two-step	-0.2032	-0.0545	0.0008	0.0964	0.0037
			(0.8798)	(0.2103)	(0.0004)	(0.5950)	(0.0168)

Table 10: Bias and MSE of estimates in CC probit and two-step Procedure when n=100 and the binary response variable y is less balanced.

Response Rate			Variable				
Cs	income		Intercept	Cs (x_2)	age (x_3)	income (x_4)	logdis (x_5)
70%	100%	CC	-0.2495	-0.0014	0.0012	0.1012	0.0169
			(0.8670)	(0.1526)	(0.0003)	(0.1429)	(0.0144)
		two-step	-0.1644	-0.0130	0.0013	0.0609	0.0100
			(0.5661)	(0.1219)	(0.0002)	(0.0921)	(0.0095)
50%	100%	CC	-0.4849	0.0085	0.0029	0.1725	0.0283
			(1.5235)	(0.2315)	(0.0004)	(0.2421)	(0.0227)
		two-step	-0.1854	-0.0006	0.0014	0.0671	0.0098
			(0.5881)	(0.1414)	(0.0002)	(0.0953)	(0.0096)
30%	100%	CC	-1.0390	0.0736	0.0078	0.3666	0.0299
			(4.3311)	(0.5424)	(0.0010)	(0.6737)	(0.0509)
		two-step	-0.2555	0.0149	0.0019	0.0895	0.0109
			(0.6670)	(0.1767)	(0.0002)	(0.1043)	(0.0100)
100%	70%	CC	-0.2733	0.0227	0.0008	0.1116	0.0152
			(0.8806)	(0.1548)	(0.0003)	(0.1482)	(0.0144)
		two-step	-0.0608	-0.0055	-0.0007	0.0495	0.0029
			(0.5076)	(0.1045)	(0.0002)	(0.1091)	(0.0093)
100%	50%	CC	-0.4316	0.0161	0.0020	0.1826	0.0143
			(1.4472)	(0.2340)	(0.0004)	(0.2471)	(0.0218)
		two-step	-0.0031	-0.0055	-0.0011	0.0237	0.0012
			(0.4937)	(0.1077)	(0.0002)	(0.1251)	(0.0094)
100%	30%	CC	-0.9668	0.0309	0.0053	0.4174	0.0266
			(4.1731)	(0.5148)	(0.0010)	(0.7323)	(0.0502)
		two-step	0.0437	0.0075	-0.0011	-0.0154	0.0017
			(0.5071)	(0.1138)	(0.0002)	(0.1587)	(0.0099)
70%	70%	CC	-0.3980	0.0150	0.0017	0.1724	0.0187
			(1.4490)	(0.2401)	(0.0004)	(0.2486)	(0.0228)
		two-step	-0.1229	0.0005	0.0004	0.0562	0.0080
			(0.5425)	(0.1258)	(0.0002)	(0.1064)	(0.0095)
60%	60%	CC	-0.7365	0.0576	0.0053	0.2509	0.0347
			(2.7233)	(0.3755)	(0.0007)	(0.4266)	(0.0365)
		two-step	-0.1216	0.0003	0.0005	0.0527	0.0076
			(0.5507)	(0.1359)	(0.0002)	(0.1104)	(0.0096)
50%	50%	CC	-1.5992	0.0991	0.0098	0.5651	0.0936
			(9.0418)	(0.8228)	(0.0019)	(1.2694)	(0.0940)
		two-step	-0.1758	-0.0023	0.0005	0.0842	0.0100
			(0.5983)	(0.1521)	(0.0002)	(0.1223)	(0.0100)

Table 11: Bias and MSE of estimates in CC probit and two-step Procedure when n=100 with higher R_p^2.

References

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica* **41**, 997 – 1016.

Bliss, C. (1934). The method of probits. *Science* **79**, 38 – 39.

Burr, D. (1988). On errors-in-variables in binary regression-berkson case. *Journal of the American Statistical Association* **83**, 739 – 743.

Chib, S. & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347 – 361.

Conniffe, D. & O′Neill, D. (2011). Efficient probit estimation with partially missing covariates. *Advances in Econometrics* **27A**, 209 – 245.

Cox, D. & Snell, E. (1981). *Applied statistics: Principles and examples.* Chapman and Hall, London.

Dagenais, M. (1973). The use of incomplete observations in multiple regression analysis: A generalized least squares approach. *Journal of Econometrics* **1**, 317 – 318.

Enders, C. (2010). *Applied missing data analysis.* The Guilford Press, New York.

Gourieroux, C. & Monfort, A. (1981). On the problem of missing data in linear models. *Review of Economic Studies* **48**, 579 – 586.

Greene, W. (2008). *Econometric analysis, 6th ed.* Prentice Hall, Upper Saddle River, NJ.

Hayashi, F. (2000). *Econometrics.* Princeton University Press, Princeton, NJ.

Heij, d. B. P. F. P. T. K., C. & van Dijk, H. K. (2004). *Econometric methods with applications in business and economics.* Oxford University Press, Oxford.

Laitila, T. (1993). A pseudo-R^2 measure for limited and qualitative dependent variable models. *Journal of Econometrics* **56**, 341 – 356.

Little, R. & Rubin, D. (1987). *Statistical analysis with missing data.* Wiley, New York.

Murphy, K. & Topel, R. (1985). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics* **3**, 88 – 97.

Schafer, J. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall, London.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data.* MIT Press, Cambridge, Massachusetts.