Agriculture and Applied Economics Association

Estimating Latent Variable Models When the Latent Variable is Observable

James K. Binkley and Luis M. Pena-Levano

DRAFT

Authors' Affiliation

James Binkley is Research Professor and Luis Pena-Levano is a PhD Candidate in the Department of Agricultural Economics at Purdue University.

Corresponding Author

James K. Binkley
Department of Agricultural Economics
Purdue University
403 West State St.
West Lafayette, IN 47907-2056
765-494-4261
E-mail: jbinkley@purdue.edu

Selected Paper prepared for presentation for the 2015 Agricultural & Applied Economics Association and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28.

ABSTRACT

Logit and probit models are designed to estimate latent variable models. However, there are cases that these estimates are used, even though the latent variable is fully observable. The most prominent examples are studies about obesity, where they calculate BMI based on two observed variables: weight and height squared. They translate BMI into a binary variable (e.g. obese or not obese) and this index is used to examine factors affecting obesity. This study determines the loss in efficiency of using logit/probit models versus the conventional OLS (e.g. with unknown variance). We also compare the marginal effects between these models. The results suggest that OLS is a more efficient than the logit/probit models when estimating the true coefficients, regardless of the multicollinearity, fit of regression and cut-off probability. Likewise, OLS provided unbiased marginal effects compared to both binary response models. It is also less likely to be biased. We can conclude, that according to our Monte Carlo simulation, when the latent variable is observable, it is better to use the continous value and regress it with respect to their explanatory variable instead of converting it into a latent variable.

Keywords: efficiency, logit, probit, BMI, bias, latent variable.

JEL codes: B23, C01, C18, C51

SECTION 1 – INTRODUCTION AND LITERATURE REVIEW

1.1 Relevance of the topic

Latent variables are variables which are not directly observed but rather inferred according to specific criteria. When a latent variable is used as a dependent variable in a regression, it is called an *index variable* (Greene 2008). More specifically, suppose we have the following regression model:

$$Y^* = X\beta + e \tag{1.a}$$

$$Y = 1, if\ Y^* > d \tag{1.b}$$

where $\boldsymbol{Y}^*$ as the unobserved dependent variable, $\boldsymbol{X}$ is the vector of independent variables, $\boldsymbol{\beta}$ is the vector of parameter estimates and $\boldsymbol{e}$ is the vector of error terms. All is known about $\boldsymbol{Y}^*$ is whether it exceeds or falls short of some threshold d. Then, the model is called as a *latent variable model* or *index function*.

The index function is typically estimated with probability models, using non-linear methods such as *logit* or *probit*, depending on the assumed distribution of the error e. Then, the goal is to measure the effect of $\boldsymbol{X}$ on the probability that the unobserved $\boldsymbol{Y}^*$ exceeds $\boldsymbol{d}$. Estimation of latent variable models is now routine. Nevertheless, the estimated coefficients of the models are not estimates of $\boldsymbol{\beta}$**,** but a standardized $\boldsymbol{\beta}$**.** Hence, aside from the sign, the estimates from this model lead to no direct interpretation (Greene 2008). But they are used to obtain *marginal effects*, which measures the effect of a small change in **X** on the probability the event occurs.

For example, if $\boldsymbol{Y}^*$ is a consumer's reservation price for a commodity (e.g. an automobile), all that is observed and recorded is whether the item was purchased. This indicates that the unseen reservation price exceeds the observed actual value. If reservation prices were observable, then the "index function" would be a standard regression model, easily estimated by OLS, with $\boldsymbol{\beta}$ being the estimate of $\boldsymbol{X}$ on a consumer's willingness to pay. Thus, the lack of information due to the fact that $\boldsymbol{Y}^*$ is an unobservable variable makes the binary models the second best solution. However, there are cases that these estimates are used, even though $\boldsymbol{Y}^*$ is fully observable.

The most prominent example is the case of the studies about **obesity**, which is an important topic in the nutrition area. BMI can be calculated as weight divided by the height squared. This variable is used to determine if a person is obese (BMI > 30) to create an dichotomous index (e.g. obese or not obese). Numerous studies have used this index to examine factors affecting obesity, which is in this case a binary variable (Gundersen et al. 2008, Fang, Ali, and Rizzo 2009, Ibrahim et al. 2014, de Mola et al. 2012).

In this case ***Y*** is "if a person is obese" and ***Y**** is BMI, and although the latent variable ***Y**** can be observed (and therefore the model could be fully estimable), it often is not estimated or used. Another example is in term of **education**, in which the surveys may contain the grades of the students (e.g. grade point averages (GPA) and even sometimes more detailed information). Nevertheless, sometimes it is preferred to group grades by a different scale (Kim 1999, Grant 2007) or in a pass/fail standard.

A natural question is why that might be case. There are some arguments that could provide some explanations:

(i) One possibility, which seems to be true in the literature on overweight, is that the focus of the study is on the discrete event (de Mola et al. 2012). Interest is usually not, for example, the effect of income on BMI, but its effect on whether someone is likely to be obese.

(ii) A similar argument can be made when studying the likelihood of passing a test based on a numerical score, or of getting a loan based on an individual's credit rating. What matters is whether the individual succeeds, not by how much they might succeed, so the success or failure becomes the focus of analysis. If the objective is to understand why students pass/fail, then interest is in the marginal effects on the probability of success, which is computed using the coefficients of the estimated probability function. Notwithstanding, a problem with this argument is that if the index function can be directly estimated this provides everything need to calculate the marginal effects in the usual way.

1.2 Contribution and objective of the study

Davidson and MacKinnon (2004) pointed out that "*It is interesting to ask how much efficiency in estimation (of a binary model) is lost by not observing the latent variable. Clearly something must be loss, since the binary variable like $\boldsymbol{y_t}$ must provide less information than a continuous variable like $\boldsymbol{y^*{}_t}$*". Nevertheless, a study that measures the loss in efficiency of using a binary variable model rather than OLS has not been previously conducted.

Thus, the purpose of this study is to examine the loss in efficiency of using logit/probit models versus the conventional OLS (e.g. with unknown variance). In order to do this, we use Monte Carlo methods to compare OLS vs binary response models. This information is useful for the interest of studies like those mentioned above, when the latent variable is observable but it is used only as a discrete variable.

For this research, we focus only in the logit/probit dichotomous version (e.g. the dependent variable only can take values of 0 and 1). Although this study addresses a specific case, the conclusions could apply also to the ordered probit/logit models (e.g. where the variable can take values of 0, 1, 2, ...).

SECTION 2 – THEORETICAL FRAMEWORK AND METHODOLOGY

2.1 The Data: Monte Carlo simulation

In this study we use Monte Carlo simulation methods to compare OLS to probit/logit under various conditions when ***Y*** is observed. We employ a program which is a combination of SAS procedures and matrix code. The main interest is the coefficients of the index function and the marginal effects, in particular, the efficiency of estimation and the extent of any bias. The methodology is described in figure 1.

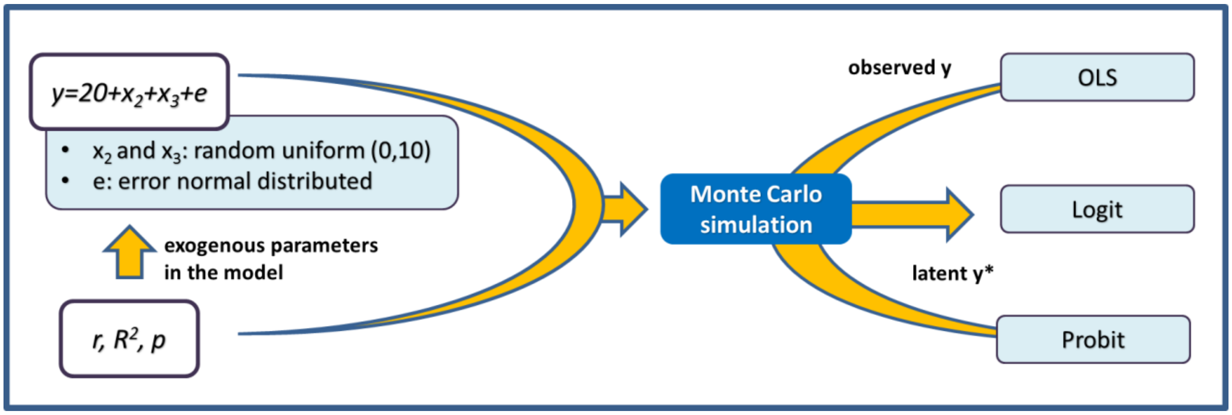
$y=20+x_2+x_3+e$
- x_2 and x_3: random uniform (0,10)
- e: error normal distributed
exogenous parameters in the model
r, R^2, p
Monte Carlo simulation
observed y
latent y*
OLS
Logit
Probit

Figure 1. Flow for the methodology of the study

The data generating process is a simple two-variable linear index function, in which we fix error variance and alter variable correlation (***r***), number of observations in a sample (***n***), the explanatory power of the model (***R²***) and the threshold of success (***c***).

We assume for this study there is not heteroskedasticity, a dependent variable (***y***) that is *observable* and two independent variables (***x₂*** and ***x₃***) with a correlation ***r*** that defines the multicollinearity between them. ***x₂*** and ***x₃*** are uniform random variables.

The total number of experiments per scenario is 50. The true model is given by the following model:

$$y^* = 20 + x_2 + x_3 + e \qquad (2)$$

where e is the error term which is normally distributed, so that probit is the theoretically appropriate estimator of the probability function.

2.2 The model scenarios

For all scenarios the error variance σ^2 is set to 1. We vary the multicollinear correlation ***r***, the ***R²*** of the model, n (number of observations per sample) and ***c*** (the threshold of success). A total of 135 scenarios per 50 experiments were obtained. 135 scenarios are the result of:

- 3 multicollinear correlations (r =0, 0.4, 0.8),
- 3 values for the fit of the model (R^2 =0.1, 0.3, 0.8),
- 3 thresholds of success for the binary response (c = 0.1, 0.25, 0.5 percentiles of ***y****) and
- 5 sample sizes (n = 30, 50, 100, 500, 1000).

2.3 The model estimations

The models to be estimated are OLS and two types of binary response models (probit/logit) following the scheme of figure 1.

(i) Binary response models

Here the latent variable is assumed to be not observed (***Y****). In this type of models, ***Y*** can only take values of 0 and 1, given a threshold ***c***:

$$Y = \begin{cases} 1\,, & if\; Y^* > c \\ 0\,, & if\; Y^* \leq c \end{cases} \qquad (3)$$

The cutoff $\boldsymbol{c_i}$ is the ith percentile of Y which varies over the experiments (i=50, 25, 10), so the probability of success ***p*** is $(1 - \boldsymbol{c_i})$ (p=0.5, 0.75 and 0.9).

The objective is to model the probability of success (***p***) given the information set **Ω**, as before. However, binary response models use a **transformation function *F(m)*** with the following properties:

$$F(-\infty) = 0, \quad F(\infty) = 1 \quad \text{and} \quad f(m) = \frac{\partial F(m)}{\partial m} > 0 \qquad (4)$$

where ***F(m)*** is a monotonically increasing function on ***m*** which lies between 0 and 1. Thus, the binary response model would have the following specification for a linear index function:

$$p = E(Y \parallel \Omega) = F(X\theta) \qquad (5)$$

thereby, $E(Y \parallel \Omega)$ would be simply a non-linear transformation ***F*** of $\boldsymbol{X\theta}$, which must lie between 0 and 1 (Greene 2008).

Thus, the regression as modeled for both cases follows, depending on the transformation of ***F*(.)**:

$$F^{-1}(y) = F^{-1}(p) = \theta_1 + \theta_2 x_2 + \theta_3 x_3 + u \qquad (6)$$

In logit/probit models, $\boldsymbol{\theta}$ does not directly estimate the effect of $\boldsymbol{X}$ on the probability[1]. The **marginal effect of the variable $\boldsymbol{X_i}$** is a nonlinear transformation $\boldsymbol{X}$, given by:

$$\frac{\partial p}{\partial X_i} = \frac{\partial F(X\theta)}{\partial X_i} = f(X\theta)\theta_i \qquad (7)$$

Furthermore, the effect on $\boldsymbol{p}$ of one of the independent variables is the greatest when $\boldsymbol{p = 0.5}$ and the least when $\boldsymbol{p}$ is either closer to 0 or 1. Two particular specifications are the most employed for ***F(.)***, which are called the **probit** and **logit** models (Davidson and MacKinnon 2004).

i.A) Probit model- For this model, the transformation function ***F*(.)** is the **cumulative standard normal distribution function** (this transformation ***F*(.)** will be denoted specifically as $\emptyset$**(.)** for a variable ***X***) given x:

$$\emptyset(x) = \int_{-\infty}^{(x)} \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}X^2)dX \qquad (8)$$

Thus, we have that the probit model is written as:

$$\emptyset^{-1}(Y) = X\theta + e, \qquad e \sim N(0,1) \qquad (9)$$

Because we only observe if ***Y*>c***, units do not matter and the variance σ^2 of ***e*** can be normalized to be equals to 1. If $\sigma^2 \neq 1$, then our coefficient $\boldsymbol{\theta}$ is not the same as $\boldsymbol{\beta}$ because it has been normalized to have a variance of 1 (Davidson and MacKinnon 2004), in other words:

$$\theta = \frac{\beta}{\sqrt{\sigma^2}} \qquad (10)$$

i.B) Logit model- The model has similar characteristics to the probit model but it is easier to deal with (Greene 2007). The transformation function ***F*(.)** is the **logistic function** $\wedge$ **(.)**:

$$\wedge (\dot{m}) = \frac{e^m}{1+e^m} \qquad (11)$$

Thus, the logit model can be written as:

$$\ln(\frac{p}{1-p}) = X\theta + e \qquad (12)$$

or, alternatively (Greene 2008):

$$p = \frac{exp(X\theta)}{1+exp(X\theta)} = \wedge (X\theta) \qquad (13)$$

ii.C) Similarities and differences between the models: According to Davidson and MacKinnon (2004), both models logit and probit tends to provide similar results. The difference between both models resides in the way that $\boldsymbol{\theta}$ is scaled.

[1] In order to calculate the true intercept (β_1), we need to adjust the intercept (β_{int}) provided by the program subtracting the threshold, in other words: $\beta_1 = \beta_{int} - c$ (for more details see Greene (2008))

For both cases, we are estimating: $\theta = \frac{\beta}{\sqrt{\sigma^2}}$ in which, for the case of the logit model, the logistic distribution has a variance of $\pi^2/3$, while for the probit model, the standard normal distribution has a variance of 1. This means that usually the estimates from the logit are larger than the probit model. This means that in order to obtain the true coefficients of the index function, we need to multiply by the $\sqrt{\pi}/3$ the coefficients of the logit.

For these models, the marginal effects are calculated as in equation (7), depending on the appropriate distribution function.

(ii) Ordinary Least Square (OLS)

For the estimation, the latent variable is observed (***y****) and assuming the variance is unknown and it has to be estimated. Thus, the model is simply calculated as:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e \tag{14}$$

in which β's are the coefficients to be estimated, and the normal errors are normally distributed and homoscedastic with an estimated variance $\hat{\sigma}^2$. Then we proceed to calculate the effects of each variable in the probability of success. First we obtain $\boldsymbol{\theta_i^{OLS}}$ of each variable $\boldsymbol{i}$:

$$\theta_i^{OLS} = \frac{\hat{\beta}}{\sqrt{\hat{\sigma}^2}} \tag{15}$$

Then, in order to compute the marginal effect of each variable ($\boldsymbol{ME_i}^{\boldsymbol{OLS}}$), we approach it using the formulation:

$$ME_i^{OLS} = g\left(x_i \theta_i^{OLS}\right) \theta_i^{OLS} \tag{16}$$

where $\boldsymbol{g(.)}$ is the probability density function of the standard normal distribution of $\boldsymbol{x_i \theta_i^{OLS}}$. The standard deviations will be obtained according to equation (16) for each of the beta coefficients.

2.4 OLS vs. binary response models

As mentioned above in (10), the estimate coefficients from the binary models (with unobservable dependent variable) are not estimates of β but of θ in which $\theta = \beta/\sqrt{\sigma^2}$. Hence, aside from the sign, the estimates from this model lead to no direct interpretation. But they are used to obtain the marginal effects.

In order to compare our models, we standardize our coefficients from the OLS (which directly estimates the index function using the observable ***Y****). Thus, we use $\hat{\beta}$ and $\hat{\sigma}^2$ from estimating the index function by OLS to obtain θ. The marginal effects can then be estimated in the usual way.

Loss in efficiency for using latent variables

As we have seen, in the case where ***Y*** is a latent variable, there is a penalty represented by the loss of efficiency due that we have less information than if we would have a continuous variable (***Y****) (Greene 2007). The OLS variance-covariance matrix for the β estimates (***VC(β)_{OLS}***)is given by:

$$VC(\beta)_{OLS} = (X^T X)^{-1} \tag{17}$$

In contrast, the variance-covariance matrix of the probit model estimates for β ($\boldsymbol{VC(\beta)_{Probit}}$)is:

$$VC(\beta)_{Probit} = (X^T \Psi(\beta) X)^{-1} \tag{18}$$

where $\Psi(\beta)$ is defined as:

$$\Psi(\beta) = \frac{f^2(X\beta)}{(F(X\beta)(1-F(X\beta))} \tag{19}$$

The largest value for $\Psi(\beta)$ is 0.6366 for the probit model which is reached at $\boldsymbol{p = 0.5}$. Hence, in the best scenario, with the largest $\boldsymbol{p}$, variance of the parameter for a probit model is**:**

$$VC(\beta)_{Probit} = \frac{VC(\beta)_{OLS}}{0.6366} = \ 1.57\, VC(\beta)_{OLS} \tag{E.20}$$

Thus, probit/logit model are much less efficient than the OLS model when using latent variables (Davidson and MacKinnon 2004). This means that in theory, if we have the possibility to observe the latent variable, it would be beneficial to use it and avoid the loss in efficiency, especially if $\boldsymbol{p}$ has values close to 0 or 1.

Therefore, since OLS uses more sample information, we would expect greater estimator efficiency. In addition, the coefficients of the index function are likely to be of interest themselves, irrespective of the primary purpose of a study.

SECTION 3- RESULTS AND DISCUSSION

Here, we provide the results from the simulations explained in section 2. Because it would be redundant to explain all the scenarios, we describe the most representative results considering that the results are similar for all the scenarios. Sub-section 3.2 reports the estimation of the coefficients and standard deviation for the models (e.g. OLS, probit and logit) for a sample of the results, whereas sub-section 3.3 discusses the marginal effects of the estimates.

3.1 Results with respect to estimation of parameters of the index function

Table 1 describes the results of one of the scenarios, which has a threshold of 0.25, a low correlation (***r***=0.4) and a high fit (***R***^{2}=0.8) and a medium sample size (***n***=100). The entries in the table are the sample means the sample standard deviation over 50 iterations of the index function coefficients. The true values of these coefficients are 20 for the coefficient and 1 for the slopes, as described in equation (2). From the table 1 we see that OLS estimates have averages very close to 1 while the probit and logit estimates exceeds 1, especially for x_2. Dividing the standard deviations by $\sqrt{50}$ to get the standard deviation of the mean, we find that the probit and logit estimates for β_2 are significantly different from 1 at p-value<0.05. Thus there appears to be some bas in these estimates. Furthermore, as expected the logit and probit estimates have larger standard deviation than the OLS results (e.g. their values are over two times higher for x_2 and x_3). This illustrates a loss in efficiency in this case.

Table 1 Estimation of the models with r=0.4, *R*^{2}=0.8, c=0.25, *n*=100

Model	Variable	Coefficient	Standard deviation	Confidence interval
LOGIT	Intercept	19.47	1.886	(18.937 , 20.003)
	x_2	1.11	0.358	(1.009 , 1.211)
	x_3	1.03	0.276	(0.952 , 1.108)
PROBIT	Intercept	19.35	1.832	(18.832 , 19.868)
	x_2	1.13	0.368	(1.026 , 1.234)
	x_3	1.04	0.258	(0.967 , 1.113)
OLS	Intercept	19.96	1.333	(19.583 , 20.337)
	x_2	1.01	0.110	(0.979 , 1.041)
	x_3	0.99	0.123	(0.955 , 1.025)

We now examine the two extreme cases to observe if there is still gain in efficiency using OLS when the dependent variable is observable. First, we present one “pessimistic case” of our scenarios: to be one of the extremes of the probability of success (***p***=0.9 or threshold ***c***=0.10), high correlation (r=0.8) and low fit to the regression (R^2=0.1) and small sample size (***n***=50). These results are displayed in table 2. Here, the standard error logit and probit estimates are again much larger than the OLS estimates.

Table 2 Estimation of a pessimistic case with r=0.8, *R*^{2}=0.1, c=0.1, *n*=50

Model	Variable	Coefficient	Standard deviation	Confidence interval
LOGIT	Intercept	19.69	1.432	(19.285 , 20.095)
	x_2	1.64	3.735	(0.584 , 2.696)
	x_3	0.86	2.760	(0.079 , 1.641)
PROBIT	Intercept	19.70	1.351	(19.318 , 20.082)
	x_2	1.63	3.596	(0.613 , 2.647)
	x_3	0.76	2.672	(0.004 , 1.516)
OLS	Intercept	20.42	2.273	(19.777 , 21.063)
	x_2	1.24	1.427	(0.836 , 1.644)
	x_3	0.75	1.363	(0.364 , 1.136)

As shown in table 2, for the three models, the standard error of the estimates for β_2 and β_3 became larger than in the previous case. As a consequence, our confidence intervals become larger and the estimates are unbiased for the three models. But again, OLS was the most efficient unbiased estimator.

We examine now one of the “optimistic” scenarios: low (or null) correlation (***r***=0), high fit to the regression (**R**^{2}=0.8), large sample size (***n***=1000) and the optimal probability of success (***p***=0.5 or threshold of ***c***=0.5). The output of the models is displayed in table 3. For this case, the standard deviations for the parameter estimates of x_2 and x_3 are much lower than before due to large sample.

Table 3 Estimation of an optimistic case with r=0, *R*^{2}=0.8, c=0.5, *n*=1000

Model	Variable	Coefficient	Standard deviation	Confidence interval
LOGIT	Intercept	20.11	0.295	(20.024 , 20.191)
	x_2	0.97	0.066	(0.956 , 0.993)
	x_3	0.98	0.065	(0.961 , 0.998)
PROBIT	Intercept	19.99	0.291	(19.907 , 20.072)
	x_2	1.00	0.065	(0.980 , 1.017)
	x_3	1.00	0.062	(0.986 , 1.022)
OLS	Intercept	20.00	0.456	(19.870 , 20.128)
	x_2	1.00	0.033	(0.986 , 1.005)
	x_3	1.00	0.033	(0.989 , 1.008)

Again, despite the magnitude of the standard deviations, the results in the optimistic scenario give similar conclusions to the two previous cases: The three models provide estimate coefficients similar to the true values. However, in this case, logit provided biased results for β_2 and β_3 (significantly different from 1 at p-value<0.05). As in the previous scenarios, OLS is more efficient than logit/probit, by about the same relative magnitude.

We will now summarize results by varying one of the parameters while we fix the other three. First, we will evaluate the correlation. We use the scenario: a threshold of 0.5, with low fit in the regression (***R***^{2}=0.3), and largest sample size (***n***=1000). What varies is the correlation between x_2 and x_3 (r=0 to 0.9) and it is represented in the x-axis. Figure 1 displays the results of the standard deviation of the coefficient estimates for x_2 (y-axis).

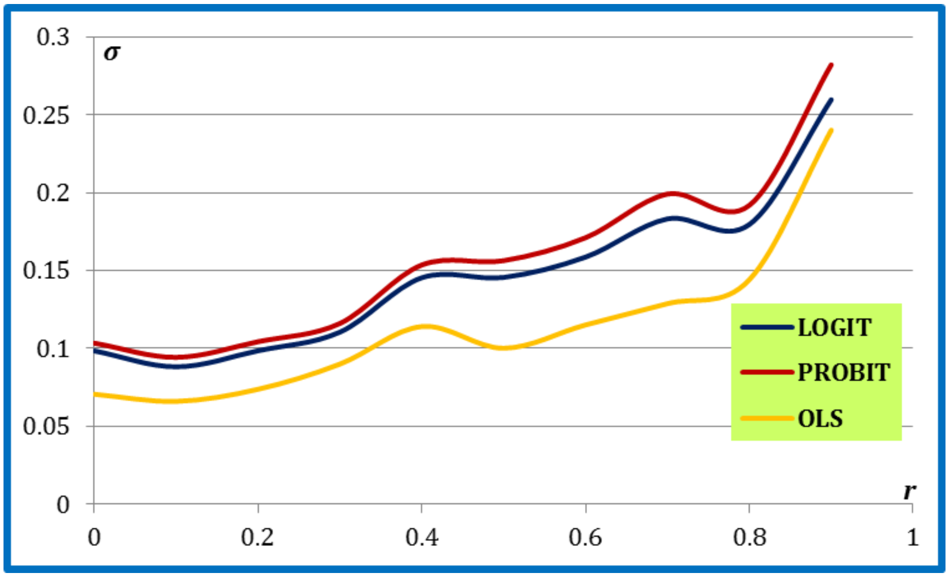

Figure 1. Standard deviation of x_2 coefficients when *c*=0.5, *R*^{2}=0.3, *n*=1000 and varying *r*

As shown in figure 1, for the three models, an increase in multicollinearity increases the standard deviation (e.g. everything else fixed). However, the standard deviation is substantially lower for the OLS results in presence of multicollinearity. On average from figure 1, OLS was approximately 34% and 40% more efficient than logit and probit, respectively. This is consistent with the theoretical results from Davidson and MacKinnon (2004) which mentions that the standard deviation of the logit/probit models would be higher than the OLS values.

On the other hand, in Figure 2, the parameter that varies is **R^2**, in this case we analize the standard deviation for the coefficients of x_3. We observe in the graph that logit/probit performs worse than OLS estimation. Also, ceteris paribus, a higher **R^2** suggest a lower standard deviation for the parameters. Because in all simulations we kept the error variance fixed at 1, increasing **R**^{2} was achieved by increasig the variation in ***X***, thus reducing the relative error variance.

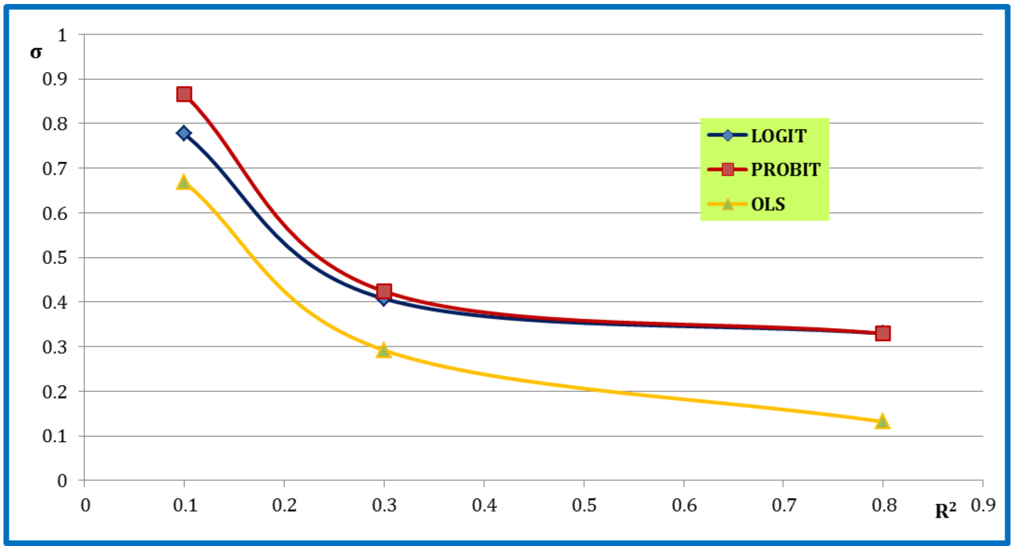

Figure 2. Standard deviation for x$_3$ coefficients when *c*=0.5, *r*=0.4, *n*=100 and varying *R*^{2}

Likewise, figure 3 displays the standard deviations for the coefficient estimates of x$_3$, when we vary the sample size ***n***, ceteris paribus. As we can observe, higher sample size decreases the variance which is consistent with theory. In this case, OLS is particularly more efficient when the sample size is small (sample ***n***=30). The difference in efficiency is much smaller when the sample size is high (***n***=1000).

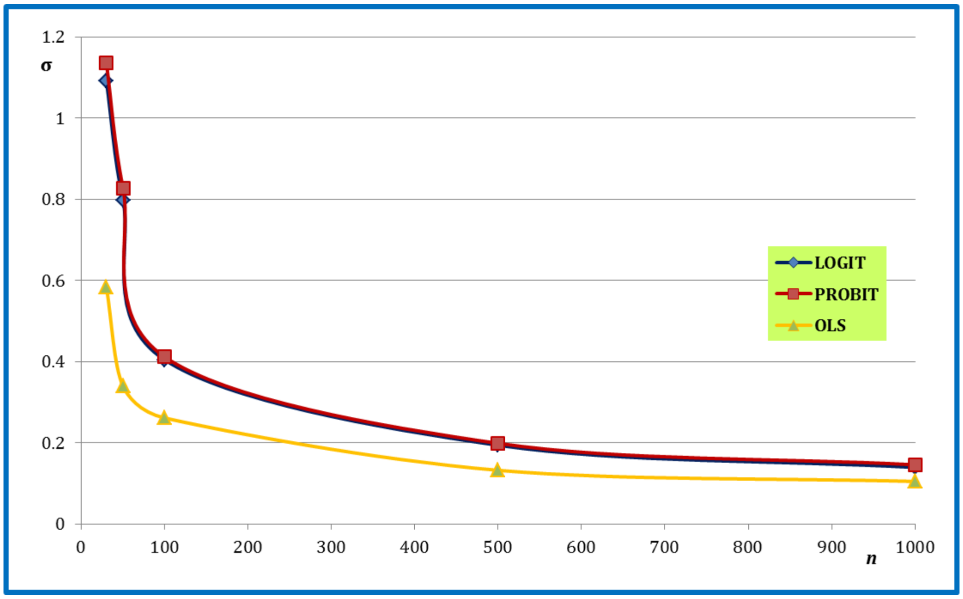

Figure 3. Standard deviation for x$_3$ coefficients when *c*=0.25, *r*=0.4, *R*^{2}=*0.3,* and varying *n*

Finally, we vary the threshold of success c, whereas we hold the other parameters constant. We analyze the standard deviation for the coefficient parameter of x_2. Figure 4 indicates that OLS performed better in terms of efficiency against both binary models. OLS shows a more constant variation independently of the threshold of success, which is not true for the logit/probit model. The difference is lower when $\boldsymbol{c}$=0.5, which is consistent with Davidson and MacKinnon (2004), which mentions that c=0.5 (which is analogous to p=0.5) provides the lowest difference in standard deviations between the two types of models.

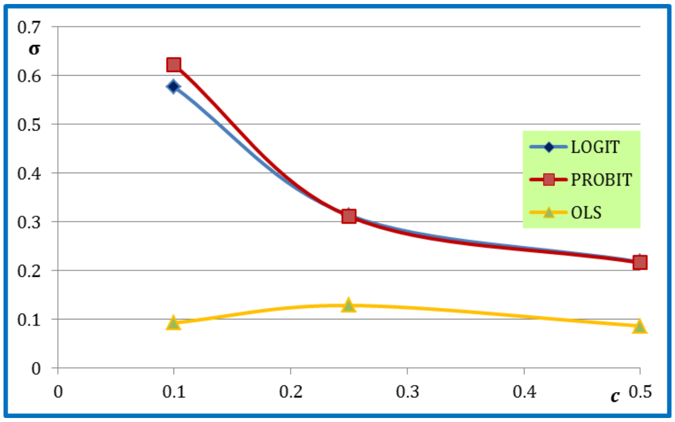

Figure 4. Standard deviation for x_3 coefficients when, *r*=0, *R^2=0.8, n*=100 and varying *c*

4.2 Results with respect to the marginal effects

We now consider marginal effects. There are two issues of importance. One is efficiency estimation, as before. The second is whether marginal effects of these models are similar to the true model. First, we consider the latter. We calculate the true marginal effects (e.g. calculated with true coefficients and the appropriate sample means). Then, we take the difference between the estimated marginal effects of each model versus the true marginal effect. We also computed the standard errors of all the difference to test for biasness.

Table 4 and 5 show two scenarios where we did these comparisons. We first note that the true marginal effects are not the same in the two cases in the tables. This is because as explained above, in order to increase $\boldsymbol{R^2}$, we increase the variation of X. Hence a "small" change in X in scenarios with different $\boldsymbol{R^2}$'s refer to different magnitude of small. This affects the marginal effects.

As we can observe in tables 4 and 5, the standard deviation of x_2 and x_3 are lower when they are estimated through OLS. The difference between these models increases substantially when the sample size is small (***n***=30). Interestingly, in both scenarios (with different $\boldsymbol{R^2}$'s), the bias is high and statistically significant (p-values<0.05) for logit and probit. In contrast, OLS provides unbiased results for the marginal effects of x_2 and x_3. In this sense, OLS is the most preferred model to be used.

Table 4 – Marginal effects and their bias when *r*=0.4, *R*^{2}=0.3, *c*=0.1, *n*=100

Model	Variable	Marginal effect	St. dev. of Marg. Effect	Bias	St. Dev. of bias	T-test of bias
LOGIT	x_2	0.10	0.044	-0.023	0.061	-2.67**
	x_3	0.10	0.041	-0.027	0.06	-3.18*
PROBIT	x_2	0.11	0.052	-0.01	0.069	-1.02
	x_3	0.11	0.053	-0.013	0.065	-1.41
OLS	x_2	0.12	0.041	-0.006	0.049	-0.87
	x_3	0.12	0.035	-0.003	0.041	-0.52

Statistical significance star codes: 0 ** 0.1 * 0.5

Table 5 – Marginal effects and their bias when *r*=0.4, *R*^{2}=0.8, *c*=0.5, *n*=30

Model	Variable	Marginal effect	St. dev. of Marg. Effect	Bias	St. Dev. of bias	T-test of bias
LOGIT	x_2	0.59	0.327	0.217	0.33	4.65**
	x_3	0.53	0.317	0.156	0.32	3.45**
PROBIT	x_2	0.56	0.323	0.188	0.327	4.07**
	x_3	0.51	0.316	0.13	0.32	2.87*
OLS	x_2	0.40	0.090	0.02	0.09	1.57
	x_3	0.39	0.070	0.009	0.069	0.92

Statistical significance star codes: 0 ** 0.1 * 0.5

Finally we show in figure 5 two different cases, when we varied the sample size, how the lines M.E. of OLS and probit become similar with larger sample size. In contrast, there was differences when the sample size was small (***n***=30 or 50). The logit values were different than the other two models, especially with low sample size. This illustrates the small sample bias of probit/logit which does not appear to affect OLS.

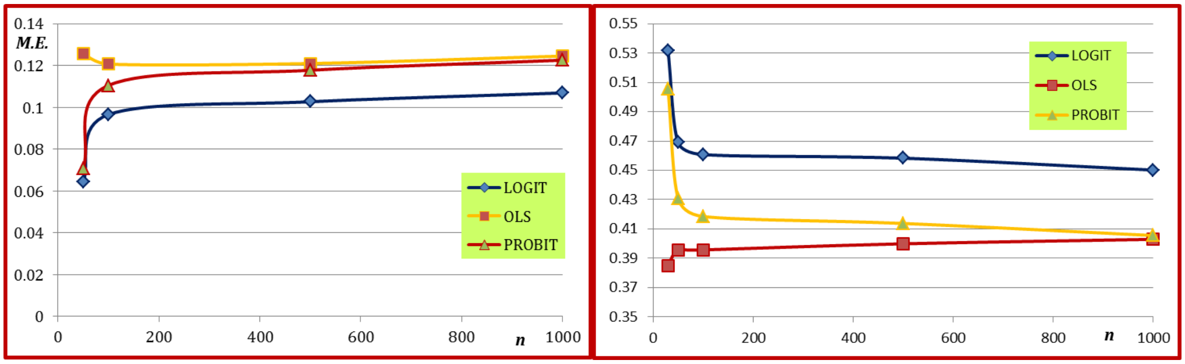

Figure 5. Marginal effects for of x_2 when *c*=0.1, *r*=0.4, R^2=0.3 (Left); and c=0.5, r=0.4, R_2=0.8 (Right) and varying *n*

Note that the graph on the right has been rescaled.

SECTION 4 – SUMMARY AND CONCLUSIONS

Logit and probit models are designed to estimate latent variable models. However, there are cases that these estimates are used, even though the latent variable is fully observable. The most prominent examples are studies about obesity, where they calculate BMI based on two observed variables: weight and height squared. They translate BMI into a binary variable (e.g. obese or not obese) and this index is used to examine factors affecting obesity.

Previous literature has not answered the question whether are still the most efficient models when the latent variable can be fully observed. Thus this study determines the loss in efficiency of using logit/probit models versus the conventional OLS (e.g. with unknown variance). We also compare the marginal effects between these models. This information is useful for the interest of food security research such as in obesity studies, specifically when the latent variable is observable but it is used only as a discrete variable.

The results suggest that OLS is more efficient than the logit/probit models when estimating the true coefficients, independently of the multicollinearity, fit of regression and cut-off probability. Likewise, OLS provided unbiased estimates in all the scenarios, which was not the case for probit/logit. These results are consistent with Davidson and MacKinnon (2004), where they mentioned that the variance of the probit models are approximately 57% higher than the OLS results when the conditions are optimal (e.g. probability of success =0.5).

In terms of marginal effects, we compared the true marginal effects with the values obtained from each model, and we found that there is biasness using logit/probit. The problem intensifies under the presence of small sample size.

We can conclude, that according to our Monte Carlo simulation, when the latent variable is observable, it is better to use the continous value and regress it with respect to their explanatory variable instead of converting it into a latent variable. This is especially the case with small sample sizes and when the probability of “success” in the population is low.

References

Bowen, William G, and Aldrich Thomas Finegan. 1969. "Economics of labor force participation."

Davidson, Russell, and James G MacKinnon. 2004. *Econometric theory and methods*. Vol. 5: Oxford University Press New York.

de Mola, Christian Loret, Timesh D Pillay, Francisco Diez-Canseco, Robert H Gilman, Liam Smeeth, and J Jaime Miranda. 2012. "Body mass index and self-perception of overweight and obesity in rural, urban and rural-to-urban migrants: PERU MIGRANT study." *PloS one* 7 (11):e50252.

Fang, Hai, Mir M Ali, and John A Rizzo. 2009. "Does smoking affect body weight and obesity in China?" *Economics & Human Biology* 7 (3):334-350.

Greene, William H. 2008. *Econometric analysis*: Granite Hill Publishers.

Gundersen, Craig, Brenda J Lohman, Steven Garasky, Susan Stewart, and Joey Eisenmann. 2008. "Food security, maternal stressors, and overweight among low-income US children: results from the National Health and Nutrition Examination Survey (1999–2002)." *Pediatrics* 122 (3):e529-e540.

Ibrahim, Chadi, Samer S El-Kamary, Jason Bailey, and Diane Marie St George. 2014. "Inaccurate weight perception is associated with extreme weight management practices in US high-school students." *Journal of pediatric gastroenterology and nutrition* 58 (3):368.