Estimated Coefficients, Information Sets, and "Biases" in Nonlinear Models

Thomas A. Mroz and Yaraslau V. Zayats

Department of Economics and the Carolina Population Center
University of North Carolina at Chapel Hill

February 2004

Abstract

Demographers often use logit and probit models when analyzing binary events. Many researchers, however, misinterpret how including or excluding additional regressors, heterogeneity corrections, and multi-level factors impact the interpretation of the estimated parameters. Such misinterpretations can result in incorrect inferences about the importance of incorporating additional features into statistical models.

In this paper we derive how estimated coefficients in probit and logit models must change when one includes or excludes explanatory "variables" that are independent of the other explanatory variables in the model. We demonstrate how coefficient estimates change when one controls or fails to control for such independent factors. Reports of "biases" in such models can often be attributed to the fact that estimates in nonlinear models depend crucially on the inclusion or exclusion of factors that are independent of those already included in the statistical model. Unlike linear regression, the set of conditioning variables plays an important role.

This is a preliminary draft and should be cited at your own risk. We wish to thank David Guilkey and Gustavo Angeles for their insightful comments as we worked on this project. Funding for this project came in part from the Measure/Evaluation project at the Carolina Population Center. The opinions expressed here are only those of the authors.

I. Introduction

Most researchers learned how to interpret regression coefficients by using the classical linear ordinary least square model. In that model, and under the standard assumption that all regressors used in the estimation are independent of the error term, each estimated regression coefficient measures the change in the expected value of the dependent variable due to a change in the regressor, holding constant all of the other observed regressors in the model. If new explanatory variables (information) becomes available, and these variables are independent of the original set of regressors, this interpretation of the coefficients as derivatives of the conditional expected value does not change when the new information has separable, additive effects.

To be precise, suppose the original regression model is given by $y_i = \beta' x_i + \varepsilon_i$. As in the classic regression model, assume $E(\varepsilon \mid x) = 0$ and ε and x are independent. In this instance the coefficient vector β equals $\partial E(y \mid x)/\partial x$. Next, suppose a new set of variables z becomes available to the researcher. These variables are independent of the original set of explanatory variables. If these additional variables were incorporated into the regression model, as in $y_i = \alpha' x_i + \gamma' z_i + \eta_i$, by following the same logic as above, the coefficient vector α equals $\partial E(y \mid x, z)/\partial x$. In what follows, we assume that η and z are independent. None of our discussion of linear models depends crucially on this assumption, but it will be a useful simplifying assumption when we discuss nonlinear models. We also assume that all functions of interest are continuous and bounded.

Since x and z are uncorrelated, and by the fact that inclusion of uncorrelated additional explanatory variables does not affect the estimates of the original variables, then $\partial E(y \mid x, z)/\partial x \equiv \partial E(y \mid x)/\partial x$ and so β≡ α. That is, the interpretation of the regression coefficient on the explanatory variables x does not depend on whether one includes or excludes the independent explanatory variables z. Of course if one was attempting to estimate the impact of x on y one would usually prefer to include z as explanatory variables, as long as z did have some explanatory power in predicting y. But this would be solely because it would be possible to obtain more accurate estimators of the effect of x in the sense of having smaller standard errors.

In most nonlinear models, however, the inclusion or exclusion of additional explanatory variables can have a substantive impact on the interpretation of estimated impacts of the original regressors, even when these additional explanatory variables are independent of all the other regressors in the model. In nonlinear models of this type, the estimates of derivatives of the expected value of outcomes depends crucially on what other variables are included in the regression model. This is little more than a statement of the fact that the expectation of a nonlinear function is seldom equal to the nonlinear function evaluated at the expectation of the explanatory variables.

To see this, suppose that y depends on a general function of x, z, and η, i.e.,

$$y_i = h(x_i, z_i, \eta_i)$$

This is just the nonlinear counterpart of the second regression model discussed above that included both x and z as explanatory variables. The expected value of y given x and z , which is the counterpart to $\alpha' x_i + \gamma' z_i$ above is $g(x_i, z_i) = \int_{-\infty}^{\infty} h(x_i, z_i, \eta) dQ(\eta)$ where Q(η) is the cumulative distribution function of the random variable η_i which, by independence, does not depend on x or z. In general g(x,z) will be a non-linear function of both x and z. The derivative of the expected value of y given x and z is $\frac{\partial E(y \mid x, z)}{\partial x} = d(x, z) = \frac{\partial g(x, z)}{\partial x}$. This derivative will usually depend on the values that x and z take; d(x,z) is a nonlinear function.

Next, suppose that information on z is not available to the researcher. In this case one cannot hold constant the value of z when describing the impact of x on y. Define the expected value of y given x as $f(x_i) = \int_{-\infty}^{\infty} g(x_i, z) dR(z)$ where R(z) is the cumulative distribution function for z. By independence R(z) does not depend on x (or η), and we treat z as a random variable because it is unobserved, just as we treated the unobserved η as a random variable. The function f(x) corresponds to the first linear regression model that includes only x as an explanatory variable. Let the derivative of this expected value be given by $\frac{\partial E(y \mid x)}{\partial x} = c(x) = \frac{\partial f(x)}{\partial x}$. This function c(x) is just an average of the d(x,z) functions across z's when g(x,z) is smooth and differentiable.

When comparing the impacts of x on y in nonlinear models with and without controlling for the impact of z, one is implicitly attempting to compare the function c(x) to the function d(x,z). Provided that both c(x) and d(x,z) are continuous and smooth and that z has continuous support, for every value of x there will always be at least one value of z, say z*(x) , that makes the two derivative equals, i.e., c(x) equals d(x,z*(x)). The values of the functions c(x) and d(x,z) will differ only because one is implicitly choosing to compare the functions at a value of z different from z*. It is only when one chooses to condition on a particular value of z that c(x) might appear to be "biased." That is, if one changes the conditioning (information) set by incorporating information about z, in nonlinear models the values of the functions c(x) and d(x,z) will usually differ. This reflects the value of new information, and should not be interpreted as a bias. c(x) will always, by definition, be an unbiased estimator of the derivative of the expected value of y given x.

Logit and probit models are two commonly used nonlinear models that display sensitivity of the parameter estimates to the inclusion or exclusion of explanatory variables that are statistically independent of the other explanatory variables used in the estimation. Many researchers, however, have misinterpreted how including or excluding additional regressors, or including or excluding heterogeneity corrections, or including or excluding multi-level factors can impact the interpretation of the estimated parameters. Such misinterpretations have given rise to several incorrect inferences about the importance of incorporating additional features into statistical models.

In this paper we derive precisely how estimated coefficients in probit and logit models can change when one includes or excludes an explanatory variable that is independent of the other explanatory variables in the model. We use these results to demonstrate how one should expect coefficient estimates to change when one controls or fails to control for these independent factors. We apply the results from our investigations to reinterpret : (1)the importance of controlling for random effects in the context of multilevel binary outcome models (Rodriguez & Goldman, 1995, 2001); (2) the suitability of a test for heterogeneity as an endogeneity test in logit models by using a conditional logit model as presented in Greene(2001); and (3) the "biases" that appear to arise when one uses Heckman-Singer heterogeneity in discrete time hazard rate models (Melino and Baker, 2000). In each of these three instances, the perceptions of "biases" that researchers might uncover could often be attributed to the failure to recognize that estimated impacts in nonlinear models depend crucially on the inclusion or exclusion of factors that are independent of those already included in the statistical model.

Section II: Probit models

In this section we derive how estimated coefficients and probability derivatives in probit models depend on the inclusion of a single explanatory variable that is statistically independent of another explanatory variable. This second explanatory variable can be used to help explain the probability of a discrete, binary event. In order for a probit model to be appropriate with both the inclusion and the exclusion of this potential explanatory variable, it is necessary for this additional explanatory variable to follow a normal distribution. If it followed some other distribution, then it would be impossible for the standard, linear index probit model to be appropriate in these two situations. The general derivation of results for this discussion for the probit model is actually quite general, and in Appendix A a similar approach is used for evaluating the consequences of including or excluding an "additional explanatory variable" in logit models.

Suppose the discrete model is of the form:

$$d_{it} = \begin{cases} 1 \text{ if } \mathrm{I}_{\mathrm{it}} \geq 0 \\ 0 \text{ otherwise} \end{cases} \qquad (1)$$

One knows that I_{it} depends on a linear function of x_1. One also believes that x_2 may enter the equation of interest linearly. Specifically, the true model is:

$$\mathrm{I}_{\mathrm{it}} = \alpha_0 + \alpha_1 \mathrm{x}_{1\mathrm{it}} + \alpha_2 \mathrm{x}_{2\mathrm{i}} + \varepsilon_{\mathrm{it}} \quad , \qquad (2)$$

where ε_{it}, x_{1it}, x_{2i} are independent of each other. We assume $\mathrm{x}_{2i} \sim \mathrm{N}(0, \sigma_{x2}^2)$ and that $\varepsilon_{it} \sim \mathrm{N}(0,1)$. These assumptions ensure that a probit model is appropriate with and without the inclusion of x_{2i} in the model. The normalization of the variance of ε_{it} to 1 is arbitrary, because only the sign of the index function I_{it} determines the outcome d_{it}; the sign of this expression does not depend on whether the entire expression is divided by any positive constant so one can pick any normalization that is convenient.

Suppose there is no information on x_{2i} at hand. Given the above model, it is the case that x_{1it} is independent of $\nu_{it} = \alpha_2 x_{2i} + \varepsilon_{it}$. This produces a 'new' probit model of the form:

$$I_{it} = \beta_0 + \beta_1 x_{1it} + \nu_{it} \qquad (3)$$

Equations (2) and (3) are identical, except for the explicit inclusion of x_{2i} in (2) and its implicit inclusion in (3). Since data is not observed for x_{2i}, and since x_{1it} is independent of ν_{it}, it seems that a probit model would be as appropriate for equation (3) where x_{2i} is not observed as it is for equation (2) where only ε_{it} is unobserved. It is, however, important to recognize that probit estimation procedures contained in standard statistical packages almost always assume, arbitratily, that the error variance is 1.0. This is true by assumption for equation (2), but it cannot be the case for equation (3) where the error term ν_{it} has variance $\alpha_2^2 \cdot \sigma_{x2}^2 + 1$.

Standard statistical procedures would produce estimates of the impact of x_{1it} on the probability that $d_{it} = 1$ under the assumption that the error term in the index function (3) has variance 1 instead of $\alpha_2^2 \cdot \sigma_{x2}^2 + 1$. We can use this fact to help derive what a probit model applied to equation (3) would estimate in terms of the parameters defined in the context of equation (2).

If one does not observe x_{2it}, then all that can be learned from the data is how the probability of the discrete event d_{it} varies with changes in x_{1it}. Using equation (2) and the fact that x_{2it} is independent of x_{1it} and ε_{it}, one can solve for this conditional probability in terms of a standard normal error term as would be imposed in a typical probit analysis. For simplicity we do not explicitly use the i and t subscripts.

Formally:

$$\text{Prob}(d = 1 \mid x_1) = \text{Prob}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon > 0 \mid x_1)$$

$$= \text{Prob}(\alpha_0 + \alpha_1 x_1 > -(\alpha_2 x_2 + \varepsilon) \mid x_1)$$

$$= \text{Prob}\left(\frac{\alpha_0 + \alpha_1 x_1}{\sqrt{\alpha_2^2 \cdot \sigma_{x2}^2 + 1}} > \frac{-(\alpha_2 x_2 + \varepsilon)}{\sqrt{\alpha_2^2 \cdot \sigma_{x2}^2 + 1}} \mid x_1 \right)$$

$$= \text{Prob}\left(\frac{\alpha_0 + \alpha_1 x_1}{\sqrt{b}} > z \right) =$$

$$= \int_{-\infty}^{\frac{\alpha_0 + \alpha_1 x_1}{\sqrt{b}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2} \, dz$$

$$= \int_{-\infty}^{\alpha_0^* + \alpha_1^* x_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \, dz \qquad (4)$$

where $b = \alpha_2^2 \cdot \sigma_{x2}^2 + 1$, is the variance of the new composite error $(\alpha_2 x_2 + \varepsilon)$.

Note that $z = \frac{-(\alpha_2 x_2 + \varepsilon)}{\sqrt{\alpha_2^2 \cdot \sigma_{x2}^2 + 1}}$ is a normal random variable with mean 0 and variance 1. It is distributed N(0,1), and it is independent of x_1. The new parameters, denoted by the original symbols in equation (2) but with asterisks, are the original parameters in equation (2) divided by the standard deviation of the new composite error, i.e., $\alpha_j^* = \frac{\alpha_j}{\sqrt{b}}$, j=1,2. Note that these transformed coefficients must be smaller than their counterparts in equation (2) whenever x_2 is informative about d.

The probability of the event d, only conditional on x_1, is given by

$$\text{Prob}(d=1 \mid x_1) = \Phi(\alpha_0^* + \alpha_1^* x_1),$$

where $\Phi(.)$ is standard normal cumulative distribution function. Almost all computer packages assume a variance one error term, so a probit analysis with x_1 as the only explanatory variable would report estimates of $\alpha_0^* = \frac{\alpha_0}{\sqrt{b}}$ and $\alpha_1^* = \frac{\alpha_1}{\sqrt{b}}$ instead of α_0 and α_1. Under the assumption that $\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 2$, and var(x_2) = var(ε) = 1, the graph of this conditional probability as only a function of x_1 is given by:

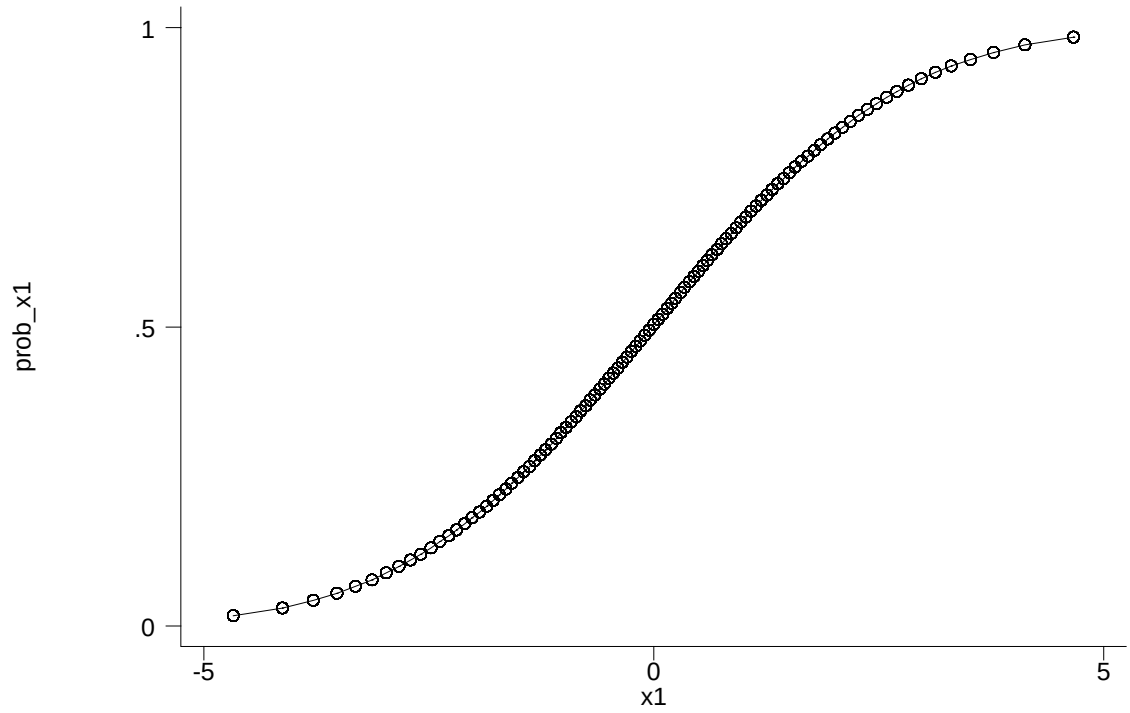

Figure 1. Prob (d=1| x_1) against x_1

This probability is clearly increasing in x_1, but it would be informative to understand quantitatively how this probability changes with x_1.

In a simple regression model, the estimates of the coefficients provide precisely this type of quantitative information. They are estimates of the derivative of the expected value of the continuous outcome with respect to a change in each explanatory variable. In nonlinear models like the probit model examined here, the estimated coefficients do not measure the derivatives of the expected value of the discrete outcome d given x, but they are related to the derivatives of this conditional expectation.

To report how changes in explanatory variables affect the expected value of the discrete outcome in the probit model, it is necessary to solve for the conditional expected value given the explanatory variables used in the estimation. Fortunately, the conditional expected value of a dummy variable (taking on only values 0 and 1) is just the probability that the dummy variable equals 1 given the explanatory variables used in the analysis. To see this note that :

$$E(d \mid x_1) = 0 \cdot \Pr ob(d = 0 \mid x_1) + 1 \cdot \Pr ob(d = 1 \mid x_1)$$
$$= \Pr ob(d = 1 \mid x_1).$$

So, in the probit model a statistic that provides information analogous to what the regression coefficients measure in a linear ordinary least squares model is just how the probability of the event d=1 varies with the explanatory variables. To solve for this piece

of information, differentiate equation (4) with respect to the explanatory variable. This yields:

$$\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1} = \frac{\partial \int_{-\infty}^{\alpha_0^*+\alpha_1^* x_1} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \mathrm{dz}}{\partial \mathrm{x}_1}$$

$$= \alpha_1^* \cdot \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(\alpha_0^*+\alpha_1^* x_1)^2}$$

$$= \alpha_1^* \cdot \phi(\alpha_0^* + \alpha_1^* x_1) \ .$$

The equality on the second line follows from the first fundamental theorem of calculus, and $\phi(.)$ is the standard normal density. Since density functions are never negative, the sign of α_1^* does indicate the sign of the derivative. However, unlike the simple linear model, the magnitude of the derivative of the probability varies by the value of x_1. Graphing this derivative, $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$, against x_1 yields:

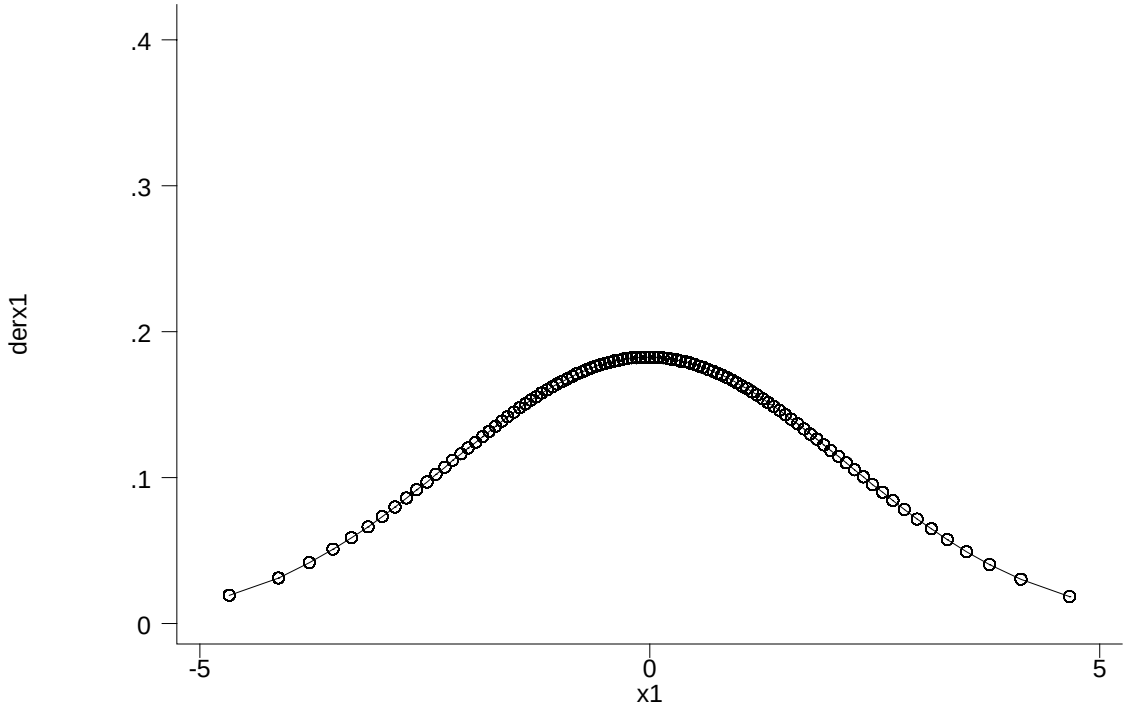

Figure 2. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ against x_1.

Since each value of x_1 implies a unique value of the Prob (d=1| x_1), we can also graph the above derivative against the probability d=1 for each value of x_1.

This yields:

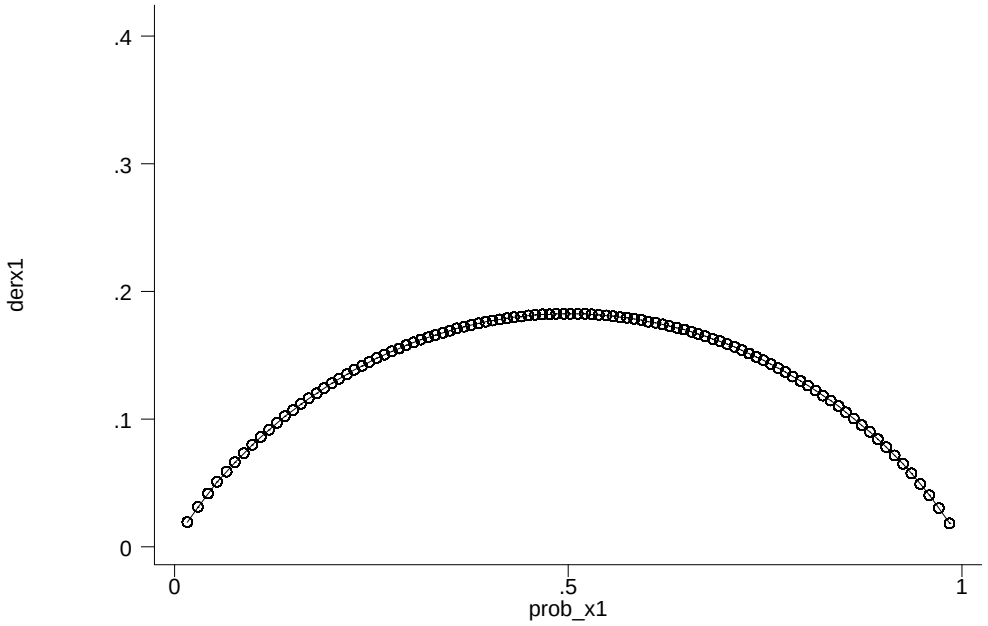

Figure 3. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ against Prob (d=1| x_1)

If there is no information available about x_2, then Figures 2 and 3 provide all of the information that one can learn about how the event {d=1} is affected by changes in the value of x_1. Probit estimators that relate the dummy variable d to only the variable x_1 provide asymptotically unbiased and consistent estimators of the coefficients α_0^* and α_1^*. These estimators will then yield consistent estimators of the derivatives of the probability that d=1 when there is no information available about x_2 (and ε).

Next, assume that information on x_2 becomes available for each observation. The researcher now can consider using the information on d, x_1 and x_2 to carry out a probit estimation of equation (2) instead of equation (3) or equation (4). Following the same steps as above, one can solve for

$$\text{Prob (d=1}\mid \mathrm{x}_1, \mathrm{x}_2) = \Phi(\alpha_0 + \alpha_1 \mathrm{x}_1 + \alpha_2 \mathrm{x}_2)$$

and

$$\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1} = \alpha_1 \cdot \phi(\alpha_0 + \alpha_1 \mathrm{x}_1 + \alpha_2 \mathrm{x}_2).$$

Note that in this instance the probability and probability derivatives correspond to the original parameters in equation (2), not the transformed parameters in equations (3) and (4). It is also important to note that the values of these statistics depend directly on the specific values of x_1 and x_2. While in a simple ordinary least squares model the impacts

of the two explanatory variables can be perfectly described by the two coefficients, in this nonlinear probit model one would need a three dimensional representation to describe all of the possible values of the impacts of the two explanatory variables.

Instead of examining all possible values of x_2, consider evaluating the probability that d=1 varies with x_1 at a few values for x_2. We also look at the average probability at each value of x_1 where we average (integrate) over all values of x_2 under the assumption that x_2 follows a standard normal distribution. Figure 4 graphs the Prob(d=1|x_1, x_2) against x_1 for three different values of x_2, namely, : -1, 0, and 1. It also contains the average Prob(d=1|x_1, x_2), where we average across values of x_2 using a standard normal distribution for this variable. This yields:

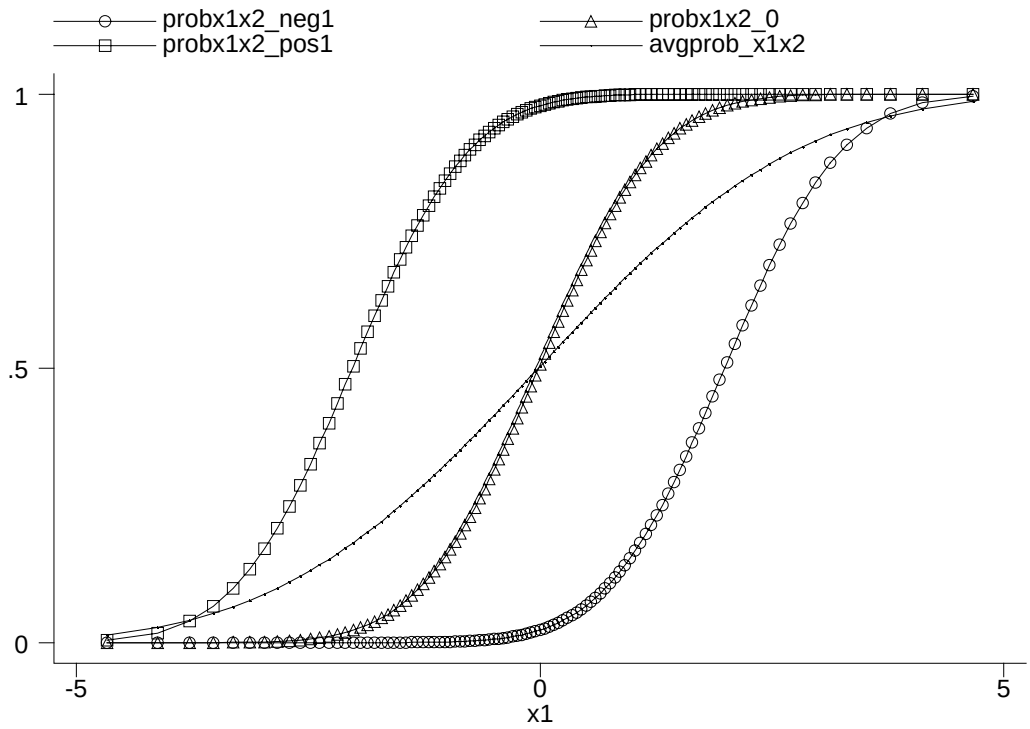

Figure 4. Prob(d=1|x_1, x_2) against x_1

Note that the curve of Prob(d=1|x_1, x_2) averaged with respect to G(x_2) (distribution of x_2) is identical to Figure 1 above.

Then we graph $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ against x_1 for the same values of x_2 (-1, 0, 1) as well as adding to the same graph $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ averaged with respect to G(x_2).

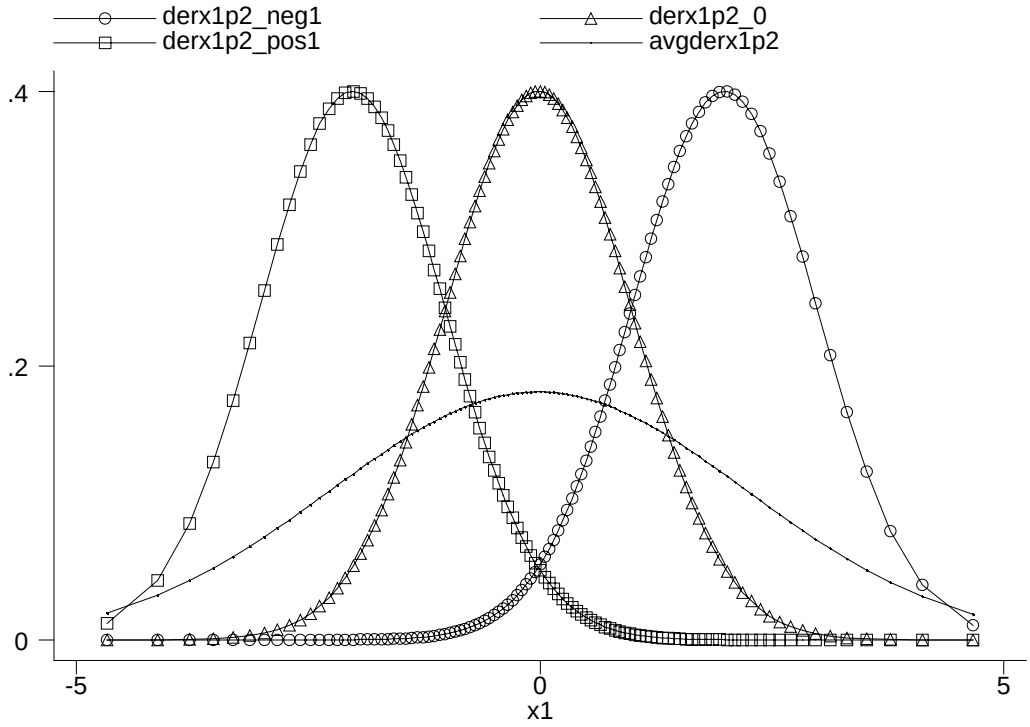

Figure 5. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_1

Note that the curve of $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ averaged with respect to $\mathrm{G}(\mathrm{x}_2)$ is identical to Figure 2.

Consider the graph of $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ against Prob(d=1|x_1, x_2):

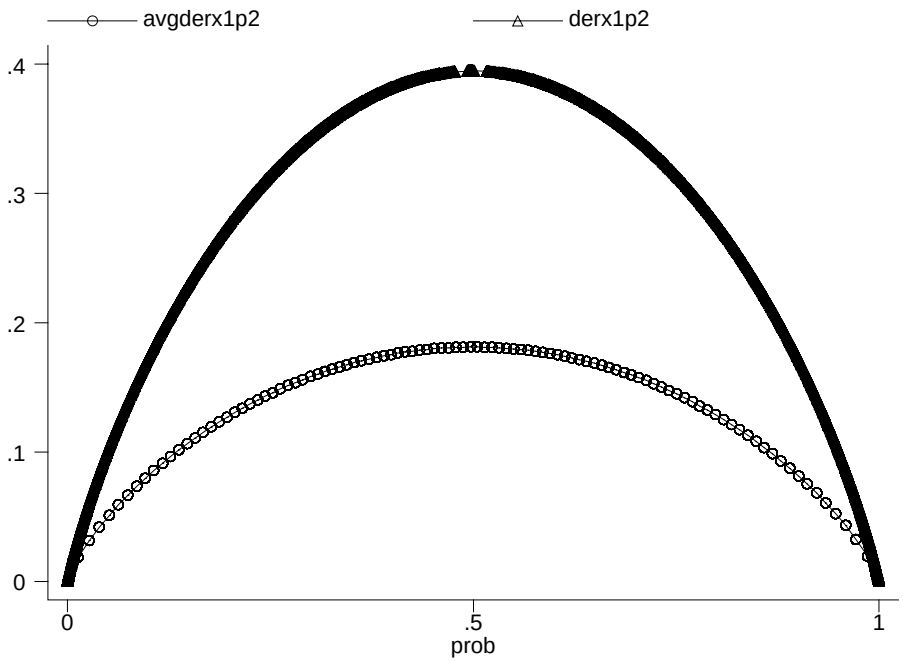

Figure 6. $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ and the averaged derivative against respective probabilities.

Comparing the two functions graphed in Figure 6, it is clear that we get different results: the marginal effects are different depending on the availability of information about x_2. One might even say that in the first case (without x_2) the derivative is wrong, and only after knowing x_2 can one obtain the correct effect of x_1 on d. Is it true? And why does this problem actually exist?

Revisit the original model:

$$d_i = \begin{cases} 1 \text{ if } I_{it} \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

Define I_{it} as

$$I_{it} = \alpha_0 + \alpha_1 x_{1it} + \alpha_2 x_{2i} + \varepsilon_{it},$$

where assume ε_{it} is normally distributed with variance equal to 1, x_{2i} is distributed normally i.e. N $(0, \sigma_{x2}^2)$, and ε_{it}, x_{1it}, x_{2i} are independent of each other. As for the actual value of x_{2i}, it might be known or it might be not, as noted already. Consider the following two cases.

1) x_{2i} is known

Then one has
$I_{it} = \alpha_0 + \alpha_1 x_{1it} + \alpha_2 x_{2i} + \varepsilon_{it}$,

In the probit model (again, let's switch for simplicity to the notation of x's as x_1 and x_2) one has:
Prob (d=1| x_1, x_2) = $\Phi(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$, where $\Phi(.)$ is standard normal distribution.
$\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1} = \alpha_1 \cdot \phi(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$, where $\phi(.)$ is the standard normal density.

2) x_{2i} is NOT known

Then, $I_{it} = \beta_0 + \beta_1 x_{1it} + \nu_{it}$, where the new error term ν_{it} is "$\alpha_2 x_{2i} + \varepsilon_{it}$" in terms of the variables from above; the standard probit normalization sets var(ν)=1 as well.

$$\text{Prob}(d=1 \mid x_1) = \text{Prob}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon > 0) = \Pr ob(\alpha_0 + \alpha_1 x_1 > -(\alpha_2 x_2 + \varepsilon)) =$$

$$= \Pr ob\left(\frac{\alpha_0 + \alpha_1 x_1}{\sqrt{\alpha_2^2 \cdot \sigma_{x2}^2 + 1}} > \frac{-(\alpha_2 x_2 + \varepsilon)}{\sqrt{\alpha_2^2 \cdot \sigma_{x2}^2 + 1}} \right) = \text{Prob}\left(\frac{\alpha_0 + \alpha_1 x_1}{\sqrt{b}} > z \right) =$$

$$= \int_{-\infty}^{\frac{\alpha_0 + \alpha_1 x_1}{\sqrt{b}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \int_{-\infty}^{\alpha_0^* + \alpha_1^* x_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz,$$

where $b = \alpha_2^2 \cdot \sigma_{x2}^2 + 1$, $z = \frac{-(\alpha_2 x_2 + \varepsilon)}{\sqrt{\alpha_2^2 \cdot \sigma_{x2}^2 + 1}}$ which is distributed N(0,1), $\alpha_0^* = \frac{\alpha_0}{\sqrt{b}}$, $\alpha_1^* = \frac{\alpha_1}{\sqrt{b}}$.

Clearly, $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1} = \alpha_1^* \cdot \phi(\alpha_0^* + \alpha_1^* x_1)$.

Solving for $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1} > \frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ for any given value of x_1 yields the ranges of x_2 where derivative of the probability of d conditional on only x_1 exceeds that when conditioning on both x_1 and x_2:

$\left(-\infty; \frac{-a - \sqrt{a^2/b + \log(b)}}{\alpha_2} \right) \cup \left(\frac{-a + \sqrt{a^2/b + \log(b)}}{\alpha_2}; +\infty \right)$, where $a = \alpha_0 + \alpha_1 x_1$, $b = \alpha_2^2 \cdot \sigma_{x2}^2 + 1$. In other words, without knowing x_2 one cannot say unambiguously whether the first marginal effect will be larger in magnitude than the second one, for a given value of x_1. Moreover, the length of the interval where the probability derivative when conditioning only on x_1 is less than that after conditioning on both x_1 and x_2 is a

nonlinear function of x_1 i.e. $\frac{2\sqrt{a^2/b+\log(b)}}{\alpha_2}$ with the minimum at $x_1 = -\frac{\alpha_0}{\alpha_1}$. To better illustrate this, Figures 7 through 11 graph both derivatives against x_2 for five different values of x_1: -3, -1, 0, 1, 3.

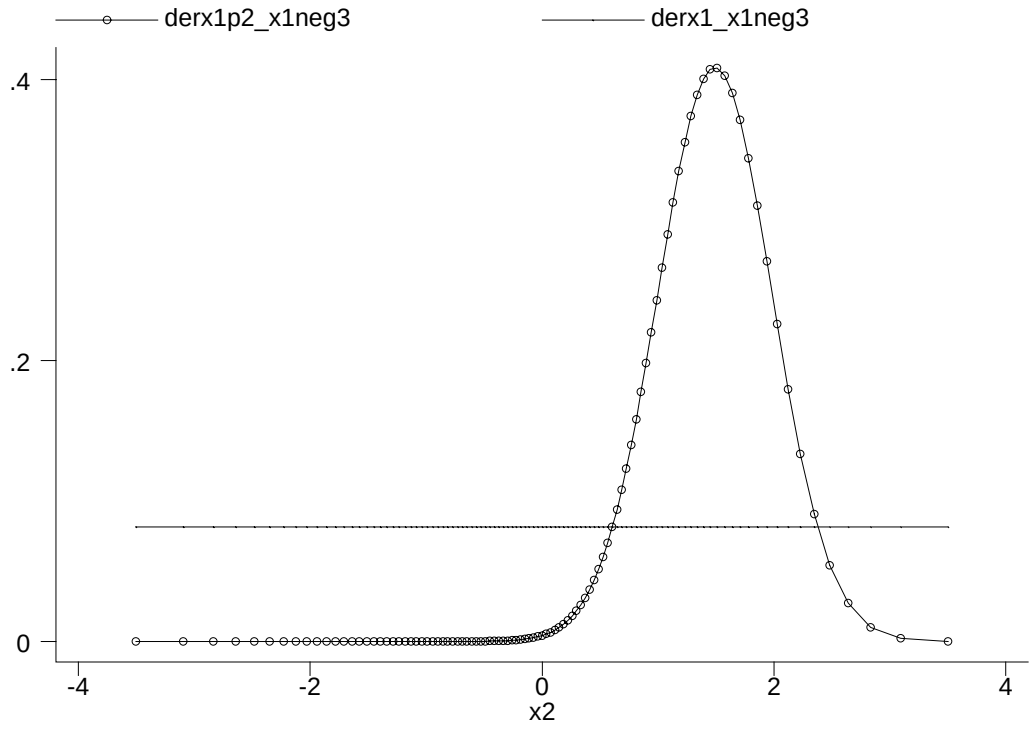

Figure 7. $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1}$ and $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ against x_2 when x_1 = -3

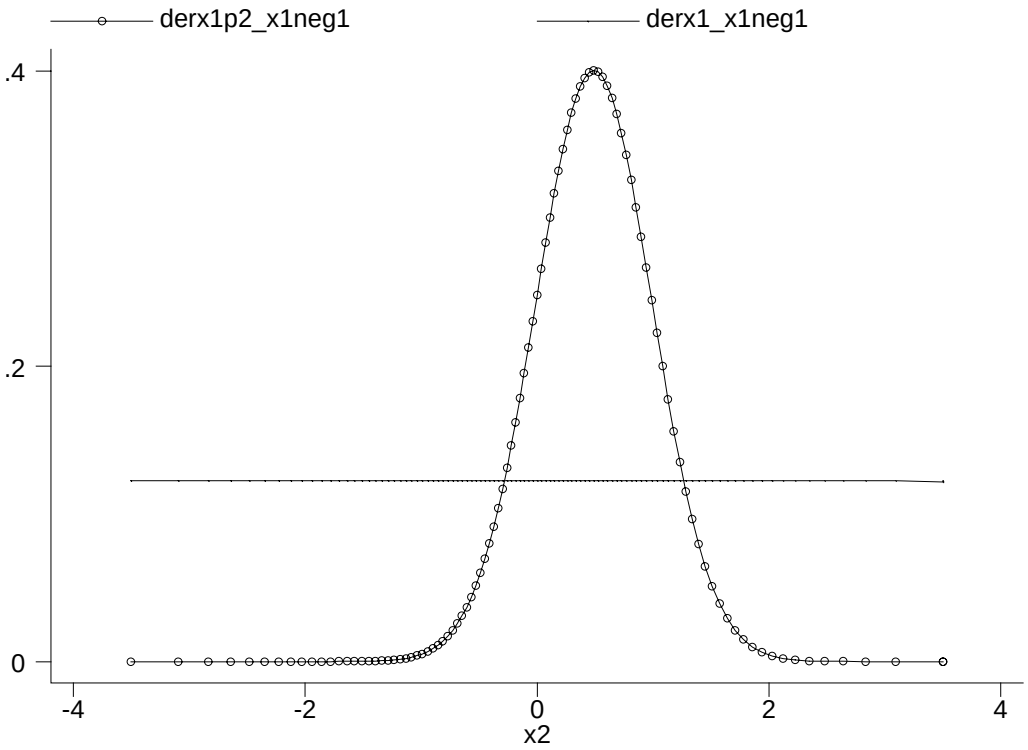

Figure 8. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = -1$

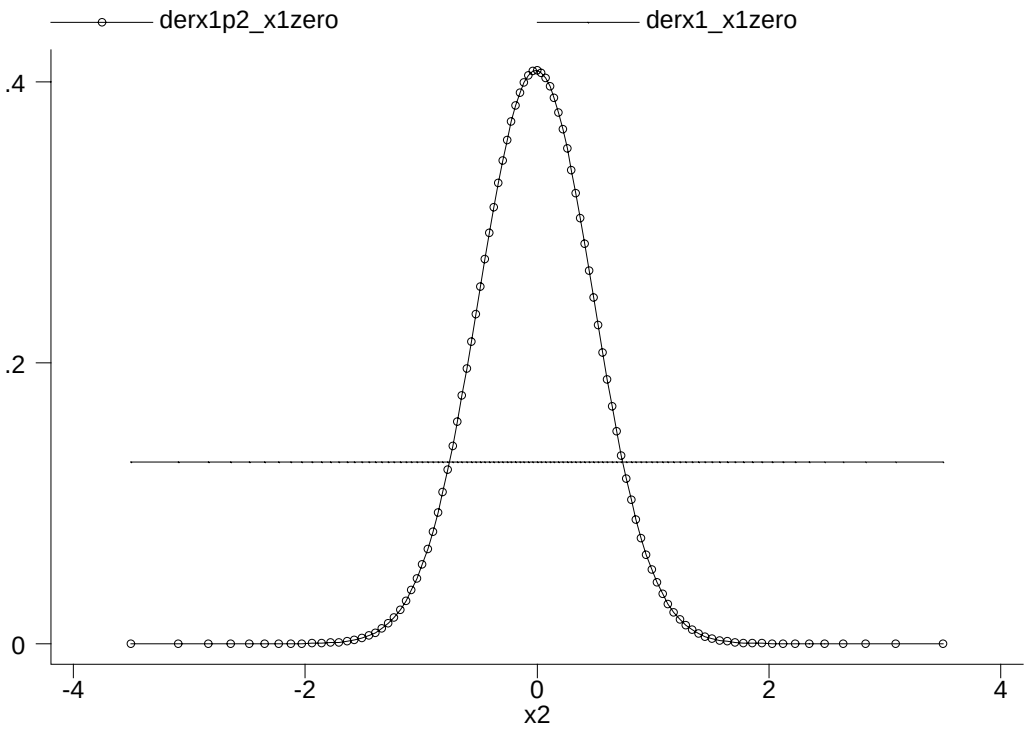

Figure 9. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = 0$

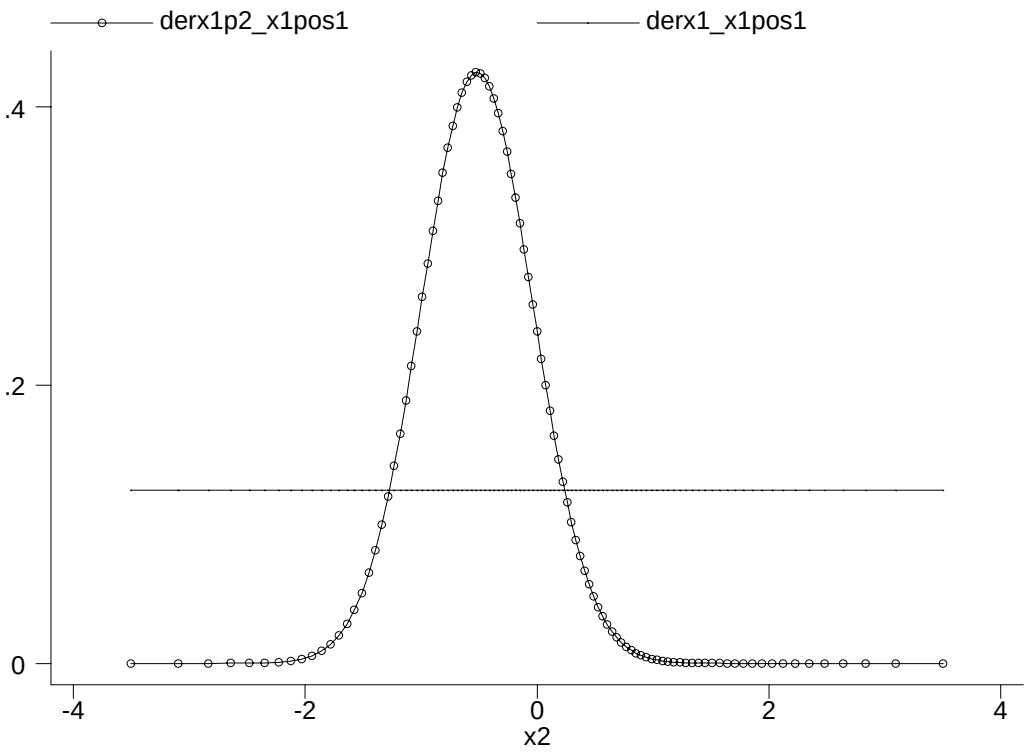

Figure 10. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = 1$

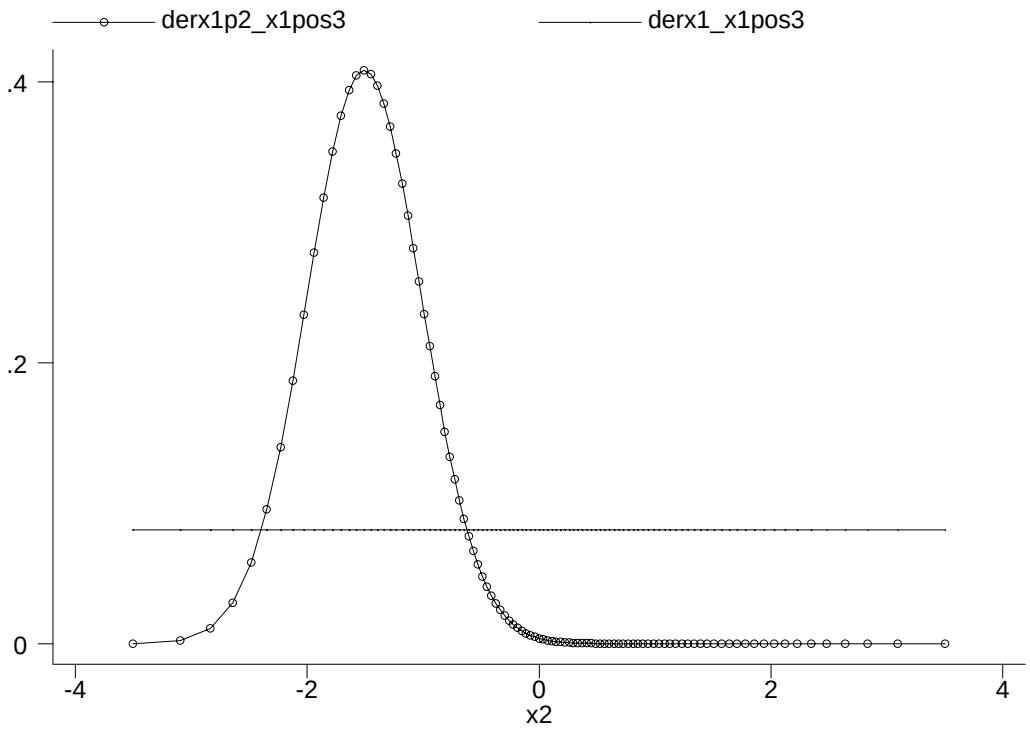

Figure 11. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = 3$

From the graphs one can see that the shortest interval (i.e. the segment of the straight line "inside" of the nonlinear derivative function) is for x_1=0. This is a 'true' minimum (in our simulations, α_0 was set to zero as was mentioned above, and so $\mathrm{x}_1 = -\frac{\alpha_0}{\alpha_1} = 0$ yields the smallest interval). Table 1 displays this a bit more compactly. Here, for each of five values for x_1, the column displays the true conditional derivative after conditioning on both x_1 and x_2 as a function of x_2. The table also highlights the values or x_2 where the conditional on x_2 and unconditional effects coincide.

For a given value of x_1 different values of x_2 yield different marginal effects. In other words having or not having information about x_2 will change one's results. This, however, does not mean that one model (e.g. when one does not have values of x_2) is necessarily biased while the other is unbiased. It is just a matter of what information the researcher considers relevant for conditioning on in her analysis. And, even if information on x_2 were available for estimation, typically one would not want to use the conditional on x_2 estimates for making inferences about situations where one cannot condition on x_2.

There is one additional implication of this analysis demonstrating that the scale of the estimated effects can often be irrelevant in one's analysis. Suppose the x_1 variable from above is, instead of being a scalar variable, a vector of two covariates determining the outcome d. Let these two elements be labeled x_{1A} and x_{1B} and let their coefficients in equation (2) be α_{1A} and α_{1B}. In many situations it is the relative magnitude of these effects that matter. For example, x_{1A} and x_{1B} might measure the intensity of two different treatments for affecting the outcome d, and one might be interested in assessing which treatment is the more cost effective. Continuing this example, suppose that an additional unit of x_{1A} costs p_A dollars and an additional unit of treatment x_{1B} costs p_B. An additional dollar spent using treatment A will result in an increase in treatment A intensity of $\frac{1}{p_A}$ units of x_{1A}. This would translate to a change in the probability index in equation (1) of $\frac{\alpha_{1A}}{p_A}$. Similarly, that one dollar spent on treatment B would increase the probability index by $\frac{\alpha_{1B}}{p_B}$. An assessment of which treatment is more cost effective would compare $\frac{\alpha_{1A}}{p_A}$ to $\frac{\alpha_{1B}}{p_B}$. If $\frac{\alpha_{1A}}{\alpha_{1B}}$ were greater than $\frac{p_A}{p_B}$, then one would conclude that treatment A was the more cost effective treatment.

If one instead estimated only the simple probit model, as in equation (3), then one would need to compare coefficients normalized to reflect the different error variance. Following the same derivations as above, the normalization factor on the coefficients of x_{1A} and x_{1B} would be identical in the simple probit model. This means that the ratio of these differently normalized coefficients would be identical to the above ratio $\frac{\alpha_{1A}}{\alpha_{1B}}$, except for sampling and estimation error. In these two models one would use exactly the same form for a test to decide which treatment was the more effective.

In many situations like this example, the most interesting interpretations of estimated covariate effects can be described by a comparison of the relative effects of covariates within a model. Such comparisons typically do not depend on the scale of the error normalization. Hence, for many policy analyses, there would be little reason to prefer the model as described in equation (2) over the model described by equation (3), except for the fact that an explicit recognition of the correlation among observations can lead to more efficient parameter estimates and more powerful tests.

Table 1

$$\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$$

	$x_1 = -3$	$x_1 = -1$	$x_1 = 0$	$x_1 = 1$	$x_1 = 3$
$x_2 = -3$	<0.0001	<0.0001.	0	0	.00443
$x_2 = -2.42323$	<0.0001	<0.0001	<0.0001	.00024	**.07253**
$x_2 = -1.17257$	<0.0001	.00148	.02550	**.16143**	.32194
$x_2 = -.63431$	.00004	.03043	**.17841**	.38480	.08911
$x_2 = -.57676$	.00007	.03925	.20510	.39426	**.07253**
$x_2 = -.17257$	.00148	**.16143**	.37587	.32194	.01175
$x_2 = 0$	.00443	.24197	.39894	.24197	.00443
$x_2 = .17257$	.01175	.32194	.37587	**.16143**	.00148
$x_2 = .57676$	**.07253**	.39426	.20510	.03925	.00007
$x_2 = .63431$	.08911	.38480	**.17841**	.03043	.00004
$x_2 = 1.17257$	.32194	**.16143**	.02550	.00148	<0.0001
$x_2 = 2.42323$	**.07253**	.00024	<0.0001	<0.0001	<0.0001
$x_2 = 3$	.00443	<0.0001	<0.0001	<0.0001	<0.0001
$\int_{-\infty}^{+\infty} \frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1} dG(x_2) = \frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1}$	**.07253**	**.16143**	**.17841**	**.16143**	**.07253**
Cutoff points (of x_2)	(.57676; 2.42323)	(-.17257; 1.17257)	(-.63431; .63431)	(-1.17257; .17257)	(-2.42323; -.57676)

Section III. Models with Logistic Distributions

A nearly identical type of analysis can be carried out for the situation where the underlying probability model is a standard, linear index logistic probability function. This analysis is presented in detail in Appendix A. The only real, substantive difference between the analyses for the probit and logit models is on the conditions for the distribution of the second explanatory variable that must be satisfied in order for the conditional on x_2 and unconditional on x_2 models to both fall in the class of linear index logistic probability functions.

For the case of the logistic function, one can again demonstrate that the derivatives are nonlinear functions of the two explanatory variables, that the derivatives depend on the information one conditions on, and that there are ranges for the second explanatory variable where the derivatives with respect to the first explanatory variable conditional on knowing the values for the second explanatory variable are smaller (or larger) than those where one does not condition on the second explanatory variable. That is, the primary lessons from the analysis of the probit case carry over to the logistic case. When interpreting estimates from multilevel logistic models, however, it is crucial to recognize that the "standard" logistic error variance is $\pi^2/3$ instead of the 1.0 typically used for the normal disturbances in probit models.

Section IV: Claims of Bias in the Literature Partially Attributable to Differences in Conditioning Sets

IV.1 Biases in Multilevel models

An influential paper by Rodriguez and Goldman (1995) has lead many researchers to conclude that, for binary outcome models in the presence of multilevel error structures, simple models that do not incorporate the multilevel error structure will yield biased estimators. The framework of the multilevel model fits neatly into the analysis presented above. In this case, the variable x_2 can be considered as the "unobserved" higher level factor giving rise to the multilevel error structure.

When one estimates a multilevel model for binary outcomes, one typically imposes the assumption that the variance of the lowest level error term follows the standard normalization assumption for the chosen binary outcome model. What this means is that comparisons of the models with and without the controls for multilevel structures implicitly are using a different variance normalization.

Consider the Monte Carlo experiments reported by Rodriguez and Goldman (1995). Part of the information on the effects of covariates in their Table 4, for effects estimated by standard logit models that ignore the multilevel structure, is displayed below in our Table 2. The implicit variance for a "standard" logistic distribution is $\pi^2/3$. The simple logit model, then, assumes that all of the variance of the error, akin to the total error as displayed in equation (2), is $\pi^2/3$. The multilevel model used to generate their

data, however, assumes that the total error variance is much larger[1]. The total, unconditional error variance in their model is $\pi^2/3+\sigma^2_{Fam}+\sigma^2_{Com}=\pi^2/3+1+1\approx 5.29$ as opposed to the implicit variance of the simple logistic model ($\pi^2/3$) that is approximately equal to 3.29.

As demonstrated above and in Appendix A, if one knows the error variance then one should be able to translate from the estimates that do not condition on the error components to those that do. This exercise is carried out in Table 2. There, we see that for each of the six average logit parameter estimates reported by Rodriguez and Goldman (1995, Table 4), the adjusted coefficients are much closer to the true parameter estimates than the unadjusted estimates[2]. These differences in estimates are due to the fact that the simple logit and the multilevel model use different information sets. The former only conditions on the observed covariates. The multilevel model, in a sense, also holds constant the unobserved error components.

Much of the evidence on bias for the logit estimators in multilevel models, then, is really evidence that the two approaches use different information sets. If the goal of the estimation is to understand the effects of changes in an observed covariate holding the error components fixed at some known level, then the multilevel model estimates would be more appropriate measures of the log-odds ratio. This would be the case where one is interested in the impacts of a covariate on one of the sampled families in one of the sampled communities. If, on the other hand, one wants to use the estimates to extrapolate to the broader population, then it would be necessary to integrate over the unobserved family and community effects as was done to derive the probability derivatives without conditioning on the second explanatory variable as in Section II. In this case the simple logit model would provide better estimates of the log-odds effect, as that model implicitly integrates over these unobserved components. Neuhaus, Kalbfleisch, and Hauck(1991) provide a more complete discussion of the interpretation of coefficient estimates in binary outcome models with correlated outcomes.

[1] From Rodriguez and Goldman (1995), p.81, "This procedure is exactly equivalent to the alternative of adding the higher level random effects to the linear predictor, calculating antilogits to obtain a conditional probability and then generating a Bernoulli random variable."

[2] If one carries out a similar exercise for Rodriguez and Goldman's (1995) Table 1 (p.83), one finds that a similar error variance adjustment, assuming that the VARCL procedure imposes an overall error variance equal to that in a standard logistic model, yields average coefficients that are quite close to the true value of 1.0 . We do not have access to the computer code for the VARCL software, so we could not evaluate whether that model does hold the overall error variance fixed. Alternatively, in these models the first order expansion of the likelihood used to estimate this model is about the point where the higher-level variances are 0, and this could be the feature that yields these parameter estimates. This appears to be the reason for Goldstein and Rasbash (1996) suggesting the use of second-order penalized quasi-likelihood in binary outcome multilevel models.

Table 2
A Re-Examination of Rodriguez and Goldman (1995)'s
Evidence on Biases in Multilevel Logit Models (From Their Table 4, p. 86)
($\sigma_{Fam} = \sigma_{Com} = 1$)

(1)	(2)	(3)	(4)	(5)
	True Parameter Value	Mean Effect Reported By Goldman and Rodriguez	Implicit Normalization, $\sqrt{\frac{\pi^2/3}{\pi^2/3+\sigma^2_{Fam}+\sigma^2_{Com}}}$	Variance Adjusted Effect, col(3)/col(4)
Guatemala Model				
Child Effect	1	0.738	0.787	0.938
Family Effect	1	0.744	0.787	0.945
Community Effect	1	0.771	0.787	0.973
Rectangular Model				
Child Effect	1	0.756	0.787	0.961
Family Effect	1	0.755	0.787	0.959
Community Effect	1	0.906	0.787	1.151

IV.2 Testing for Endogeneity with Conditional Logit Model Estimators

The conditional logit estimator has been suggested as an approach to use to test for the endogeneity of explanatory variables in a discrete outcome, panel data model. See, for example, Cecchetti's (1986) application of the test and the testing approach presented in Greene (2001, p.841). Chamberlain (1983) provides detailed information about the "fixed-effect" conditional logit estimator. The test considers whether the coefficients on the time varying explanatory variables change significantly when one uses a conditional logit estimation approach instead of a simple logit model applied to all the data. In general, it is implemented like a Hausman (1978) test for model misspecification.

If the source of heterogeneity is independent of all the explanatory variables, then one is in a situation described by comparing equations (2) and (3). Equation (2) would correspond to the model conditioning on the individual persistent effect, while equation (3) would correspond to the standard logit estimator. Both the conditional logit and the standard logit assume a standard logistic distribution for their remaining error terms, and so a direct comparison of the estimates obtained from these two models should indicate a difference in levels of the estimated coefficients whenever there is a heterogeneity. The presence of heterogeneity, however, does not by itself indicate that any of the exogenous variables are endogenous in the sense that they are correlated with the error components.

Consider the following panel data model which is a simple extension of that presented in equation (2)

$$I_{it} = \alpha_0 + \alpha_1 x_{1it} + \alpha_2 x_{2it} + \eta_i + \varepsilon_{it}, \quad d_{it} = 1 \Leftrightarrow I_{it} > 0$$

where the observed explanatory variables x_1 and x_2 are independent of the error components ε and η. We assume, as in the multilevel formulation, that ε and η are independent, that ε follows a standard logistic distribution, and that the sum of the two error components follows a general logistic distribution. If one conditions out the observation i specific effect η, then the resulting disturbances, ε, follow a standard logistic distribution.

Since the observed explanatory variables are independent of the composite error term, one can use simple logit estimator to obtain consistent estimates of the effects of x_1 and x_2 after adjusting for the variance normalization. Let this estimation model be

$$I_{it} = \alpha_0^* + \alpha_1^* x_{1it} + \alpha_2^* x_{2it} + \nu_{it}^* \quad ,$$

$$where \quad \alpha_j^* = \alpha_j \sqrt{\frac{\pi^2/3}{Var(\eta) + \pi^2/3}}$$

$$and \quad Var(\nu_{it}^*) = \pi^2/3$$

Whenever the Var(η) is not equal to zero, and the conditional logit model and the simple logistic model will estimate different coefficients. In the former case these correspond to a logit model holding constant the η, while in the latter case the estimates correspond to a logistic model after integrating out the observation specific components η. Therefore, finding that the coefficients α_j^* do not equal α_j is not compelling evidence that the explanatory variables x_1 and x_2 are endogenous in the sense of the explanatory variables being correlated with the error term. The coefficients might differ only because one is implicitly conditioning on different information sets. Such a test, however, is a test for the presence of heterogeneity.

There is, however, a simple test that can be used to test for the exogeneity of x_1 and x_2 with respect to the component η. In particular, since the error variance normalization is identical for all coefficients, then under the null hypothesis that the x_1 and x_2 are independent of η the ratio of any two coefficients from the standard logit model should equal the ratio of the coefficients corresponding to the same two variables in the conditional logit model. For example, the test statistic might be whether the difference $\frac{\hat{\alpha}_1^*}{\hat{\alpha}_2^*} - \frac{\hat{\alpha}_1}{\hat{\alpha}_2}$ is significantly different from zero. If it were, then this would be evidence that at least one of the explanatory variables is not independent of the error component η. The standard error of the estimated difference could easily be calculated from a bootstrap procedure where one samples on observations i (not independently on the (i,t) pairs) to reflect the fact that the composite error terms could be correlated across t for any observation.

IV.3 Biases in the Estimation of Hazard models with Heckman-Singer Semi-parametric Heterogeneity Controls

A recent paper by Baker and Melino (2000) explores the use of Heckman-Singer unobserved heterogeneity in discrete time hazard models. The main point of their study deals with the number of support one should use to approximate the distribution of unobserved heterogeneity in such hazard models. They find that as one uses additional discrete points of support to approximate the heterogeneity distribution that the estimated coefficient on the observed covariates tend to rise to levels above those specified in the data generating process. The primary recommendation from their analysis is that one should be quite conservative in adding additional points of support to the estimated heterogeneity distribution.

A key point ignored in their analysis is that the variance of the estimated heterogeneity distribution, given that they are only approximating the true model, need not closely resemble the true heterogeneity distribution. Consequently, their interpretations and comparisons of absolute levels of the coefficients on observed covariates could be comparing coefficients estimated with quite different error normalizations. This is precisely what the discussion in Section II suggests can lead to incorrect inferences. The model Baker and Melino (2000) examine, however, is much more complex that those discussed above. This is due to the fact that a hazard rate model with unobserved heterogeneity, by definition, implies a data generating process with potentially severe sample selection biases. Nonetheless, it is informative to examine whether using simple normalizations to control for the differences in the error variances in their discrete outcome models might alter the conclusions of their study.

For the most part one cannot interpret the simulation results reported by Baker and Melino (2000) in a way that allows one to address whether adjustments for different error variances across estimation approaches might alter the main conclusions of their analysis. The information they present in their Table 1, for a single replication for one data generating mechanism for three different sample sizes, however, does contain enough information to assess the sensitivity of their results to adjustments for error variances for that one replication. While not definitive, the results are quite suggestive.

Table 3 contains information from Baker and Melino's (2000) Table 1 plus some calculations made using their estimates of the estimated heterogeneity distributions for these three samples. The adjustment factor used in Table 3 is the ratio of the true total error variance to that estimated by the discrete heterogeneity model, i.e., $\left(1+\pi^2/3\right)\Big/\left(\hat{Var}+\pi^2/3\right)$. A comparison, for this one replication, of the estimated β coefficients and the adjusted coefficients displayed in the last column indicates that the adjusted coefficients are much closer to the true parameter value than the estimates reported by Baker and Melino (2000)[3]. In particular for those models using three and four points of support, the resulting "biases" fall by over fifty percent after recognizing that

[3] Baker and Melino (2000) present two additional pieces of information that support this inference of their reported biases being due to their failure to recognize the differences in error variance. First, they find that the models with the "excessively" large coefficients predict the underlying hazard rates quite well (pp. 382-5). Second, they find that the predicted expected values of the duration conditional on covariate values are more accurate when one uses higher number of points of support (p.385-6). It is typically these quantities that are of interest to researchers.

the absolute level of the coefficients can only convey information about interesting magnitudes after normalizing on the error variance[4]. If this single observation is not an aberration, their finding of large biases from using nonparametric heterogeneity corrections might be called into question. Additionally, their recommendation to penalize severely models with higher numbers of points of support when choosing the controls for unobserved heterogeneity might lead to underspecified empirical models. For example, Fenton and Gallant (1996) and Mroz (1999) each presents Monte Carlo evidence that criteria like the Akaike Information Criteria can lead to underspecified models and inaccurate predictions when one is attempting to approximate an unknown distribution. The Hannan-Quinn Information Criterion suggested by Baker and Melino (2000) will tend to be more conservative in adding points of support whenever the sample size is greater than 55. The difficulties Baker and Melino (2000) find for estimating the distribution of unobserved heterogeneity in the presence of unknown forms of duration dependence, however, does suggest that single spell duration models can provide very little information about unobserved heterogeneity.

[4] Ideally one would compare relative magnitudes of effects as noted above. But since Baker and Melino's(2000) Monte Carlo only used a single observed explanatory variable, we are forced to compare estimated coefficients to estimated error standard deviations.

Table 3
A Re-Evaluation of Baker and Melino's (2000) Estimates from Their Table 1

Number of Points of Support	True β	Estimated β	Estimated Heterogeneity Variance	Adjustment to β for comparison with True DGP	Adjusted β
Sample size: 500					
1	1.00	0.731	0.00	1.142	0.834
2	1.00	0.740	27.863	0.371	0.275
3 (inferior)*	1.00	1.079	24.258	0.395	0.426
3	1.00	1.615	3.796	0.778	1.257
4	1.00	1.737	4.326	0.751	1.304
Sample Size: 1000					
1	1.00	0.665	0.00	1.142	0.759
2	1.00	0.671	4.228	0.755	0.506
3	1.00	0.976	0.810	1.023	0.998
4	1.00	1.333	2.287	0.834	1.112
Sample Size: 5000					
1	1.00	0.670	0.00	1.142	0.765
2	1.00	1.066	1.085	0.991	1.056
3	1.00	1.216	1.942	0.906	1.101
4	1.00	1.269	2.219	0.885	1.123

* This estimation achieved a lower likelihood function value than the exact same specification's in the next row.

Section V: Conclusions

The interpretation of coefficient estimates from binary outcome models with unobserved error components is complicated by the fact that different statistical model specifications implicitly alter the interpretation of the coefficients on explanatory variables. The results presented above indicate that one needs to be somewhat cautious when comparing estimates from different estimation procedures, even those within roughly the same family of estimators. A simple extension of these results suggests that one should exercise a large degree of caution when interpreting the coefficient estimates from binary outcome models obtained from different samples even if they use identical estimation procedures. The reason for all of these results is that interpretation of the regression coefficients in binary outcome models as indicators of true magnitudes crucially depends on arbitrary normalizations that are used in the estimation of binary outcome models.

While direct comparisons and interpretations of the coefficients from different estimation approaches for binary outcomes can be quite problematic, nearly all interesting magnitudes related to probabilities and impacts of covariates on predicted probabilities can be compared across different models. In fact, in many important instances comparisons of relative effects of covariates are directly comparable across estimation approaches. If researchers would focus on policy-relevant numerical quantities, rather than conveniently chosen "effects," nearly all of the interpretation problems we have highlighted in this paper would vanish. This, in fact, is the main conclusion of this paper.

There would still be problems related to the sensitivity of predicted probabilities and probability derivatives to variations in the conditioning sets used to define the probabilities. Many issues that are nearly completely irrelevant in standard linear regression models can have key impacts on the magnitudes of effects estimated with nonlinear models. For the cases illustrated here, whether one should or should not condition on unobserved heterogeneity in the interpretation of the effects of observed covariates can be crucially important for inferences from nonlinear models. Additionally, as opposed to the intuition researchers have developed for interpreting linear models, the effect of an observed covariate in a nonlinear model usually depends critically on the exact values taken on by <u>all</u> explanatory variables. Precise definitions of the substantive effects researchers want to uncover, however, would usually resolve many of these issues. There certainly could be differences in opinion among researchers about what are the correct effects to consider. However, we could learn much more if researchers would state the reasons why one should prefer inferences about one type of specific effect to another. An arbitrary label of "bias" applied to an estimation procedure because its implicit normalization does not correspond to just one particular and often arbitrary definition of a type of effect, without a statement of why that particular type of effect is of key importance, could lead to many invalid inferences.

Appendix A: The Logistic Model

Suppose the discrete model is of the form:

$$d_i = \begin{cases} 1 \text{ if } I_{it} \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

One knows that I_{it} depends on a linear function of x_1. One also believes that x_2 may enter the equation of interest linearly. Specifically, the true model is:

$$I_{it} = \alpha_0 + \alpha_1 x_{1it} + \alpha_2 x_{2i} + \varepsilon_{it} \qquad (5)$$

where ε_{it}, x_{1it}, x_{2i} are independent of each other. We assume ε_{it} is distributed logistically with mean 0 and variance $\frac{\pi^2}{3}$; and x_{2i} is distributed such that $\frac{\varepsilon_{it} + \alpha_2 x_{2i}}{2}$ is distributed logistically (0, $\frac{\pi^2}{3}$). These assumptions ensure that a logit model is appropriate with and without the inclusion of x_{2i} in the model.

Suppose there is no information on x_{2i} at hand. Given the above model, it is the case that x_{1it} is independent of $\nu_{it} = \alpha_2 x_{2i} + \varepsilon_{it}$. This produces a 'new' logit model of the form:

$$I_{it} = \beta_0 + \beta_1 x_{1it} + \nu_{it} \qquad (6)$$

If one does not observe x_{2it}, then all that can be learned from the data is how the probability of the discrete event d_{it} varies with changes in x_{1it}. Using equation (5) and the fact that x_{2it} is independent of x_{1it} and ε_{it}, one can solve for this conditional probability in terms of a logistical error term as would be imposed in a typical logit analysis. For simplicity we do not explicitly use the i and t subscripts:

Formally:

$$\text{Prob}(d = 1 \mid x_1) = \text{Prob}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon > 0) = \Pr ob(\alpha_0 + \alpha_1 x_1 > -(\alpha_2 x_2 + \varepsilon)) =$$

$$= \Pr ob\left(\frac{\alpha_0 + \alpha_1 x_1}{2} > \frac{-(\alpha_2 x_2 + \varepsilon)}{2} \right) = \text{Prob}\left(\frac{\alpha_0 + \alpha_1 x_1}{2} > z \right) = \Lambda(\frac{\alpha_0}{2} + \frac{\alpha_1}{2} x_1) = \Lambda(\alpha_0^* + \alpha_1^* x_1),$$

where $z = \frac{-(\alpha_2 x_2 + \varepsilon)}{2}$ which is distributed $\Lambda(0, \frac{\pi^2}{3})$, $\alpha_0^* = \frac{\alpha_0}{2}$, $\alpha_1^* = \frac{\alpha_1}{2}$, $\Lambda(.)$ is logistic distribution. Under the assumption that $\alpha_0 = 0$, $\alpha_1 = 1$, $\alpha_2 = 2$ the graph of this conditional probability as only a function of x_1 is given by:

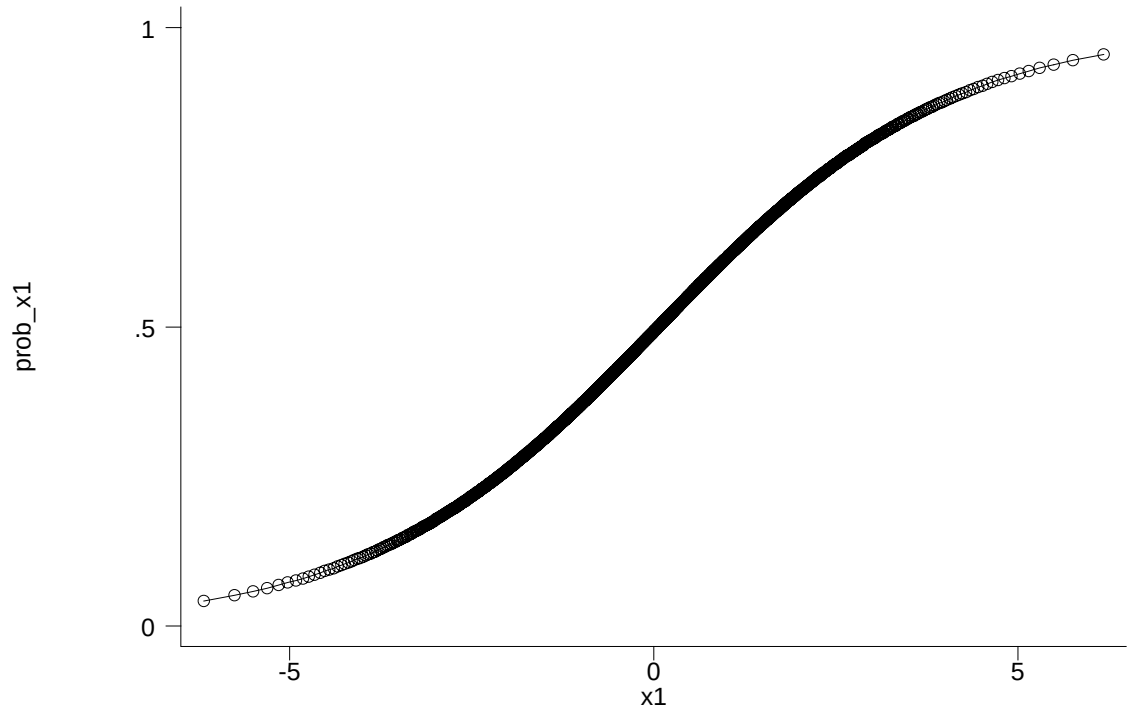

Figure 12. Prob (d=1| x_1) against x_1

This probability is clearly increasing in x_1, but it would be informative to understand quantitatively how this probability changes with x_1.

From the derivations above it follows that $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1} = \alpha_1^* \frac{\exp(\alpha_0^* + \alpha_1^* x_1)}{(1+\exp(\alpha_0^* + \alpha_1^* x_1))^2}$.

Graphing this derivative, $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1}$, against x_1 yields:

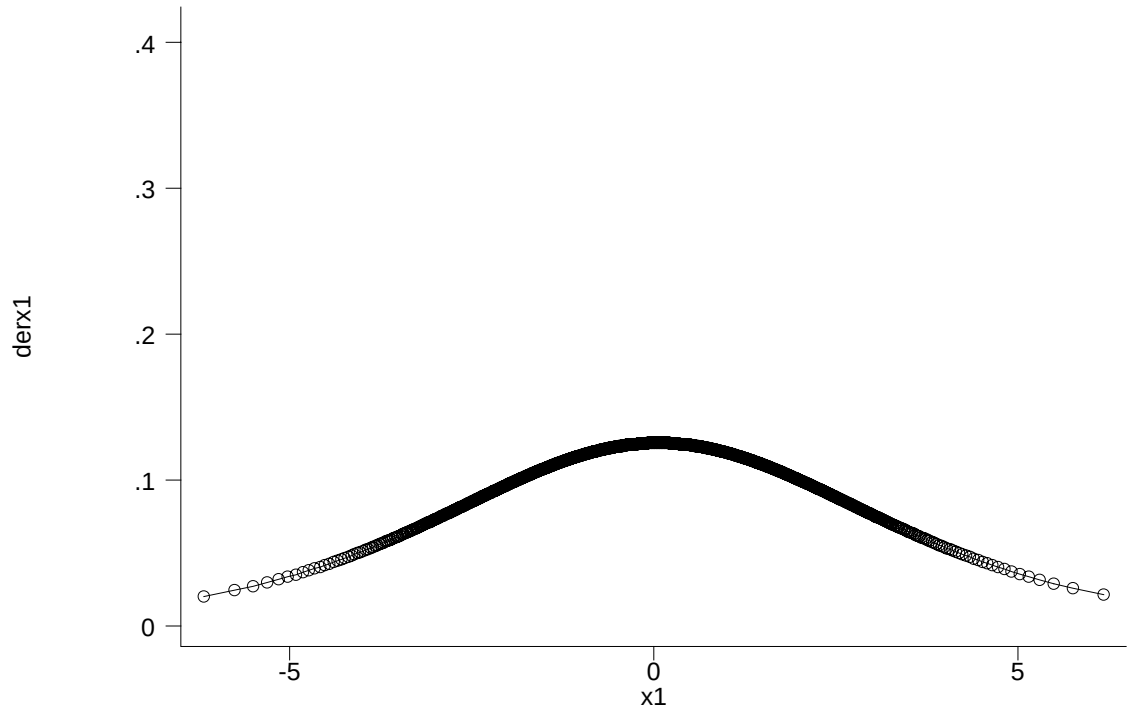

Figure 13. $\frac{\partial \Pr ob(d = 1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ against x_1.

Since each value of x_1 implies a unique value of the Prob (d=1| x_1), we can also graph the above derivative against the probability d=1 for each value of x_1. This yields:

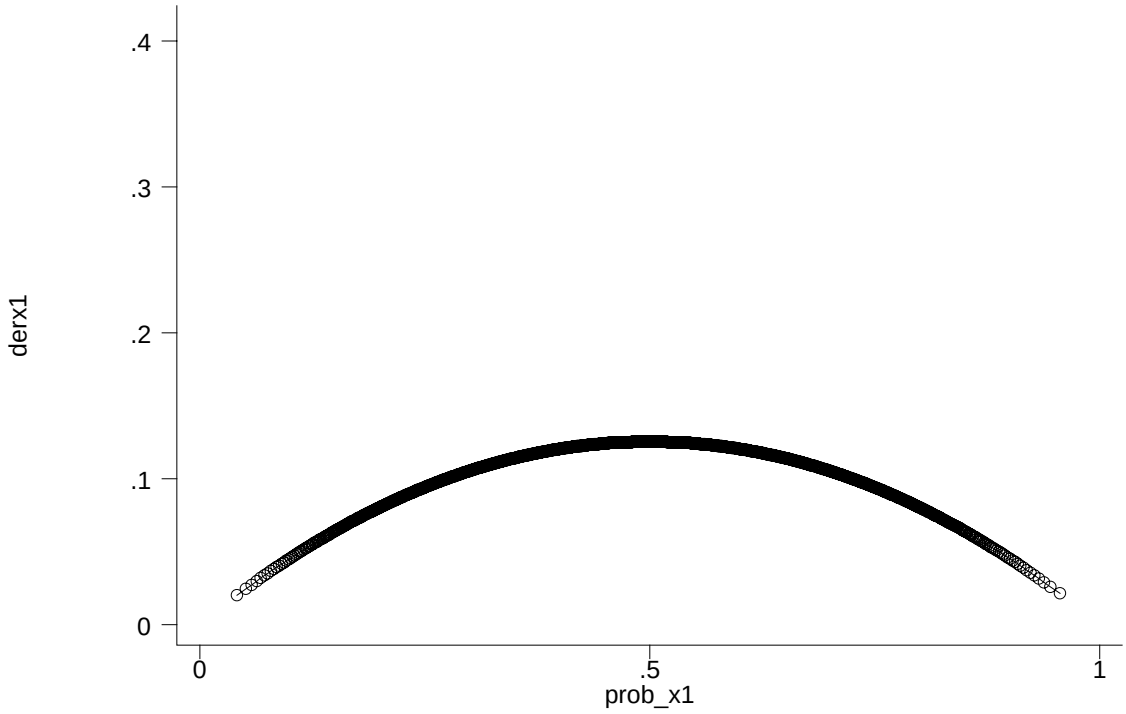

Figure 14. $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1}$ against Prob (d=1| x_1)

If there is no information available about x_2, then Figures 13 and 14 provide all of the information that one can learn about how the event {d=1} is affected by changes in the value of x_1. Logit estimators that relate the dummy variable d to only the variable x_1 provide asymptotically unbiased and consistent estimators of the coefficients α_0^* and α_1^*. These estimators will then yield consistent estimators of the derivatives of the probability that d=1 when there is no information available about x_2.

Next, assume that information on x_2 becomes available for each observation. The researcher now can consider using the information on d, x_1 and x_2 to carry out a logit estimation of equation (5) instead of equation (6). Following the same steps as above, one can solve for

Prob (d=1| x_1, x_2)= $\Lambda(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$ and

$$\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1} = \alpha_1 \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)}{(1+\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2))^2}.$$

Note that in this instance the probability and probability derivatives correspond to the original parameters in equation (5), not the transformed parameters in equation (6). It is also important to note that the values of these statistics depend directly on the specific

values of x_1 and x_2. While in a simple ordinary least squares model the impacts of the two explanatory variables can be perfectly described by the two coefficients, in this nonlinear probit model one would need a three dimensional representation to describe all of the possible values of the impacts of the two explanatory variables.
Instead of examining all possible values of x_2, consider evaluating the probability that d=1 varies with x_1 at a few values for x_2. We also look at the average probability at each value of x_1 where we average over all values of x_2.
Figure 15 graphs the Prob(d=1|x_1, x_2) against x_1 for three different values of x_2, namely, : -1, 0, and 1. It also contains the average Prob(d=1|x_1, x_2), where we average across values of x_2.

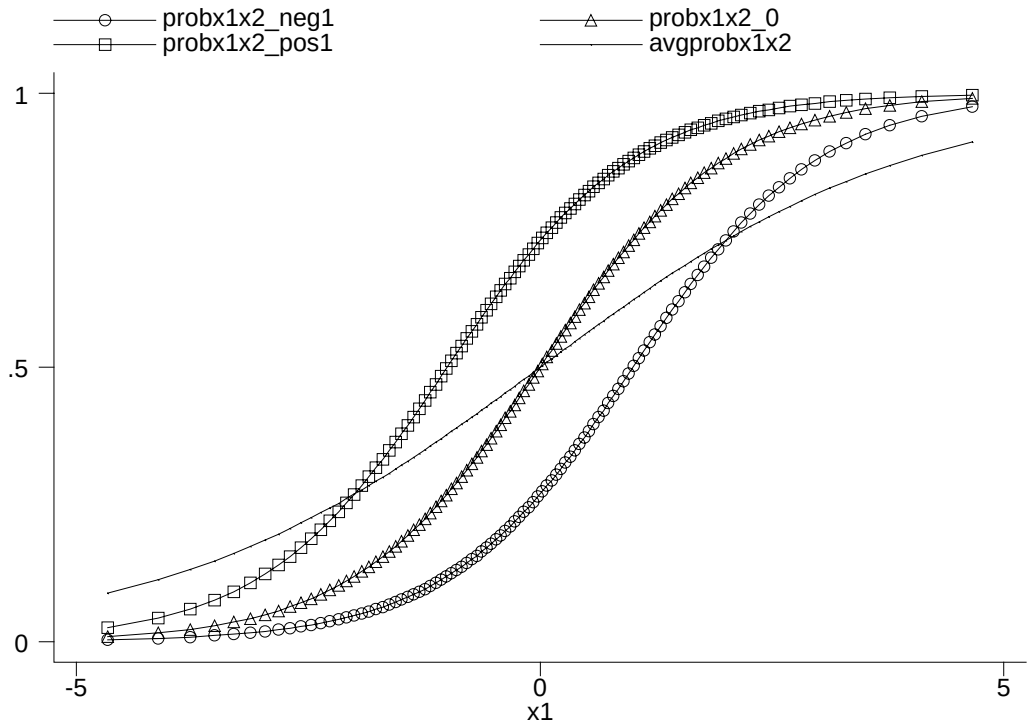

Figure 15. Prob(d=1|x_1, x_2) against x_1

Note that the curve of Prob(d=1|x_1, x_2) averaged with respect to G(x_2) is identical to Figure 12 above.
Then we graph $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ against x_1 for the same values of x_2 (-1, 0, 1) as well as adding to the same graph $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ averaged with respect to G(x_2).

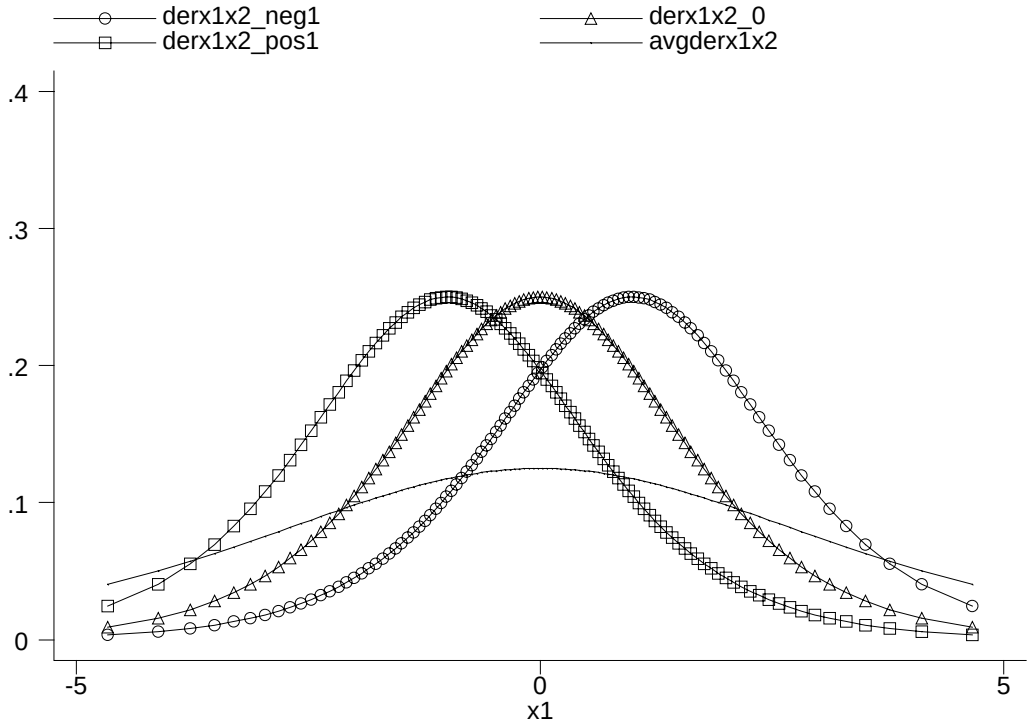

Figure 16. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_1

Note that the curve of $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ averaged with respect to $\mathrm{G}(\mathrm{x}_2)$ is identical to Figure 13.

Consider the graph of $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ against Prob(d=1|x_1, x_2):

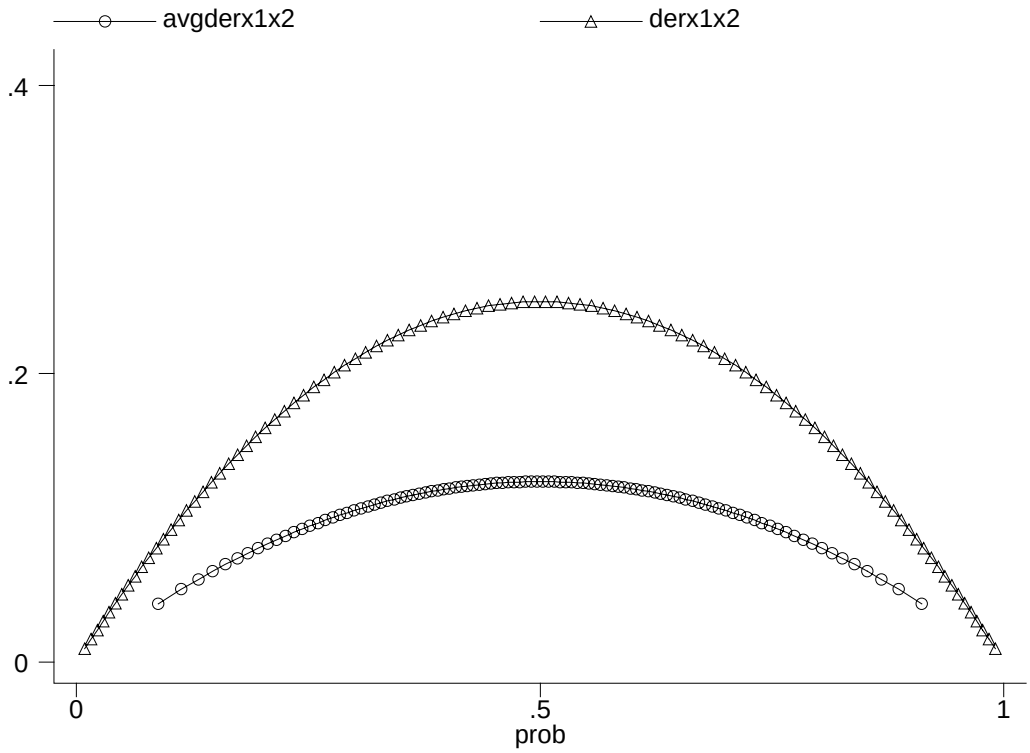

Figure 17. $\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ and the averaged derivative against respective probabilities.

Comparing the two functions graphed in Figure 17, it is clear that we get different results: the marginal effects are different depending on the availability of information about x_2. One might even say that in the first case (without x_2) the derivative is wrong, and only after knowing x_2 can one obtain the correct effect of x_1 on d. Is it true? And why does this problem actually exist?

Revisit the original model:

$$d_i = \begin{cases} 1 \text{ if } I_{it} \geq 0 \\ 0 \text{ otherwise} \end{cases}$$

Define I_{it} as

$$I_{it} = \alpha_0 + \alpha_1 x_{1it} + \alpha_2 x_{2i} + \varepsilon_{it},$$

where ε_{it}, x_{1it}, x_{2i} are independent of each other, ε_{it} is distributed logistically with mean 0 and variance $\frac{\pi^2}{3}$, x_{2i} is distributed such that $\frac{\varepsilon_{it} + \alpha_2 x_{2i}}{2}$ is distributed logistically (0, $\frac{\pi^2}{3}$).

As for the actual value of x_{2i}, it might be known or it might be not, as noted already. Consider the following two cases.

1) x_{2i} is known

Then one has

$$I_{it} = \alpha_0 + \alpha_1 x_{1it} + \alpha_2 x_{2i} + \varepsilon_{it},$$

In logit model (again, let's switch for simplicity to notation of x's as x_1 and x_2) one has:

Prob (d=1| x_1, x_2) = $\Lambda(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$, where $\Lambda(.)$ is logistic distribution.

$$\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1} = \alpha_1 \frac{\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)}{(1+\exp(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2))^2}.$$

2) x_{2i} is NOT known

Then, $I_{it} = \beta_0 + \beta_1 x_{1it} + \nu_{it}$, where the new error term ν_{it} is "$\alpha_2 x_{2i} + \varepsilon_{it}$" in terms of the variables from above.

$$\text{Prob}(d=1 \mid x_1) = \text{Prob}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon > 0) = \Pr ob(\alpha_0 + \alpha_1 x_1 > -(\alpha_2 x_2 + \varepsilon)) =$$

$$= \Pr ob\left(\frac{\alpha_0 + \alpha_1 x_1}{2} > \frac{-(\alpha_2 x_2 + \varepsilon)}{2}\right) = \text{Prob}\left(\frac{\alpha_0 + \alpha_1 x_1}{2} > z\right) = \Lambda(\frac{\alpha_0}{2} + \frac{\alpha_1}{2} x_1) = \Lambda(\alpha_0^* + \alpha_1^* x_1),$$

where $z = \frac{-(\alpha_2 x_2 + \varepsilon)}{2}$ which is distributed $\Lambda(0, \frac{\pi^2}{3})$, $\alpha_0^* = \frac{\alpha_0}{2}$, $\alpha_1^* = \frac{\alpha_1}{2}$, $\Lambda(.)$ is logistic distribution.

Clearly, $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1} = \alpha_1^* \frac{\exp(\alpha_0^* + \alpha_1^* x_1)}{(1+\exp(\alpha_0^* + \alpha_1^* x_1))^2}$.

Solving for $\frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1} > \frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$ for any given value of x_1 yields the ranges of x_2 where derivative of the probability of d conditional on only x_1 exceeds

that when conditioning on both x_1 and x_2:

$$\left(-\infty;\frac{\log\left(1+a+a^2-(1+a)\times\sqrt{1+a^2}\right)-\log(a)-\alpha_0-\alpha_1 x_1}{\alpha_2}\right)\cup$$

$$\cup\left(\frac{\log\left(1+a+a^2+(1+a)\times\sqrt{1+a^2}\right)-\log(a)-\alpha_0-\alpha_1 x_1}{\alpha_2};+\infty\right)$$

where $a=\exp(\alpha_0^*+\alpha_1^* x_1)$. In other words, without knowing x_2 one cannot say unambiguously whether the first marginal effect will be bigger than the second one, for a given value of x_1. Moreover, the length of the interval where the probability derivative when conditioning only on x_1 is less than that after conditioning on both x_1 and x_2 is a nonlinear function of x_1 i.e.

$\frac{\log\left(1+a+a^2+(1+a)\times\sqrt{1+a^2}\right)-\log\left(1+a+a^2-(1+a)\times\sqrt{1+a^2}\right)}{\alpha_2}$ with the minimum at $x_1=0$. To better illustrate this, Figures 18 through 22 graph both derivatives against x_2 for five different values of x_1: -3, -1, 0, 1, 3.

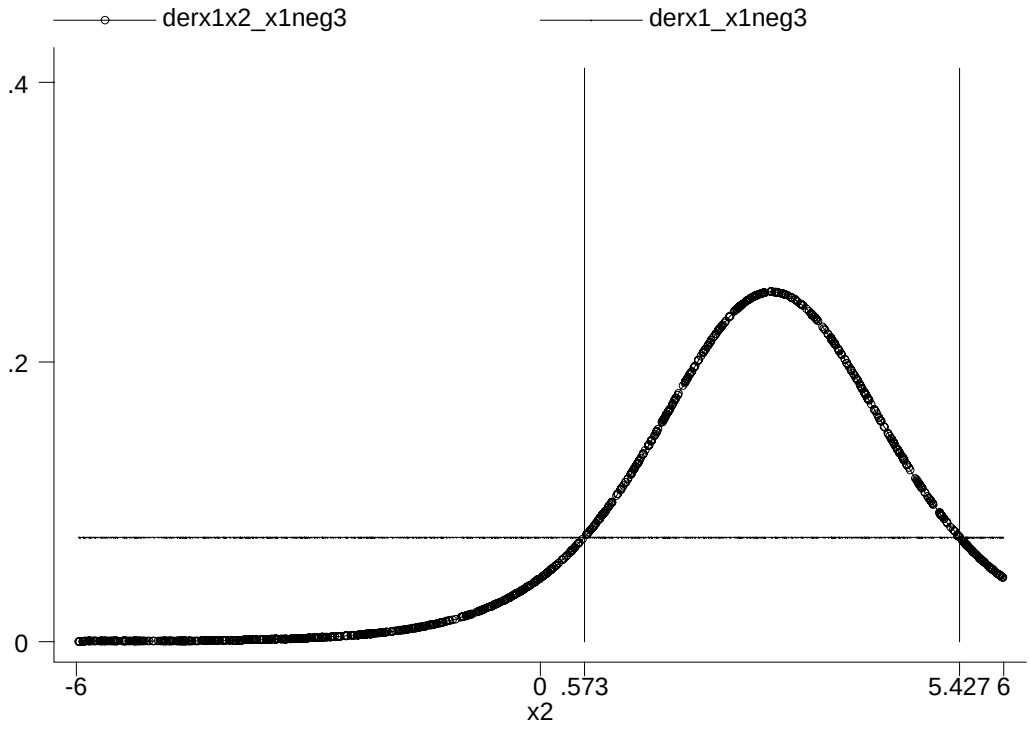

Figure 18. $\frac{\partial \Pr ob(d=1\mid x_1)}{\partial x_1}$ and $\frac{\partial \Pr ob(d=1\mid x_1,x_2)}{\partial x_1}$ against x_2 when x_1= -3

Vertical lines at values of x_2 `.57331 and 5.42669` indicate that the averaged derivative and specific values of the derivative are equal.

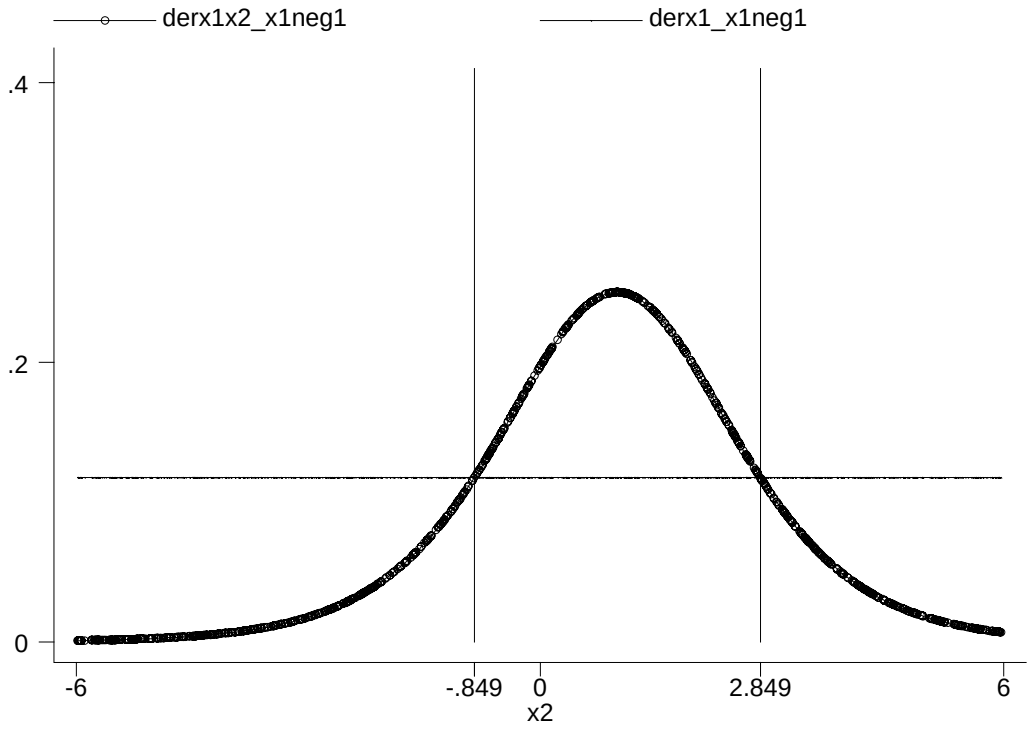

Figure 19. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = -1$

Vertical lines at values of x_2 `-.84894 and 2.84894` indicate that the averaged derivative and specific values of the derivative are equal.

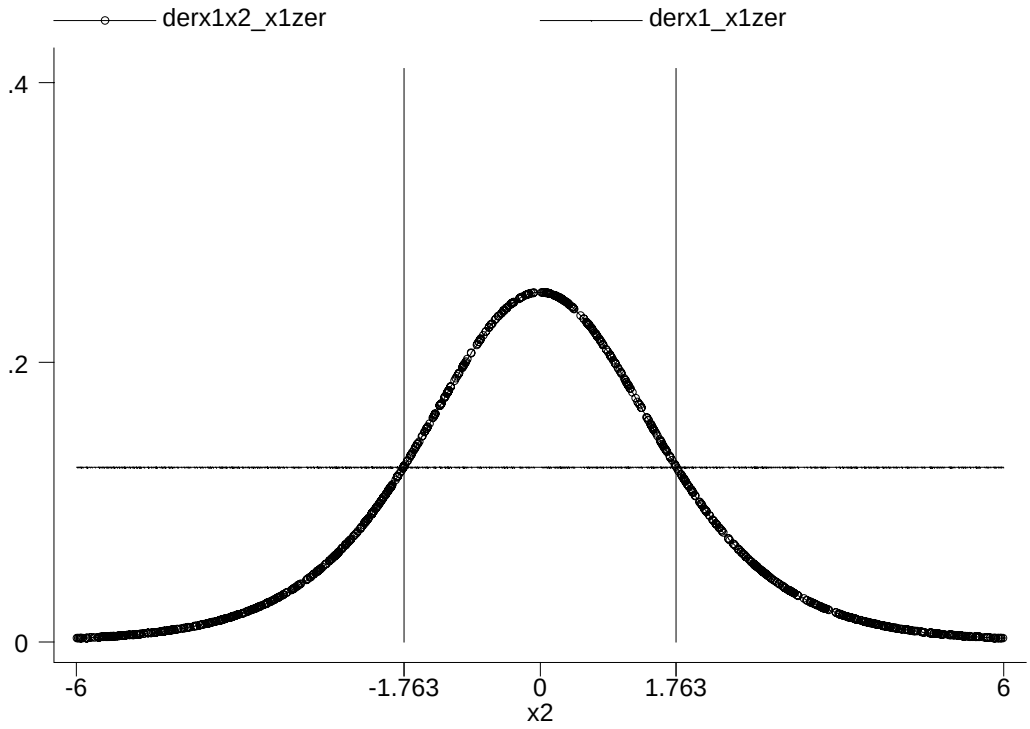

Figure 20. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = 0$

Vertical lines at values of x_2 `-1.76275 and 1.76275` indicate that the averaged derivative and specific values of the derivative are equal.

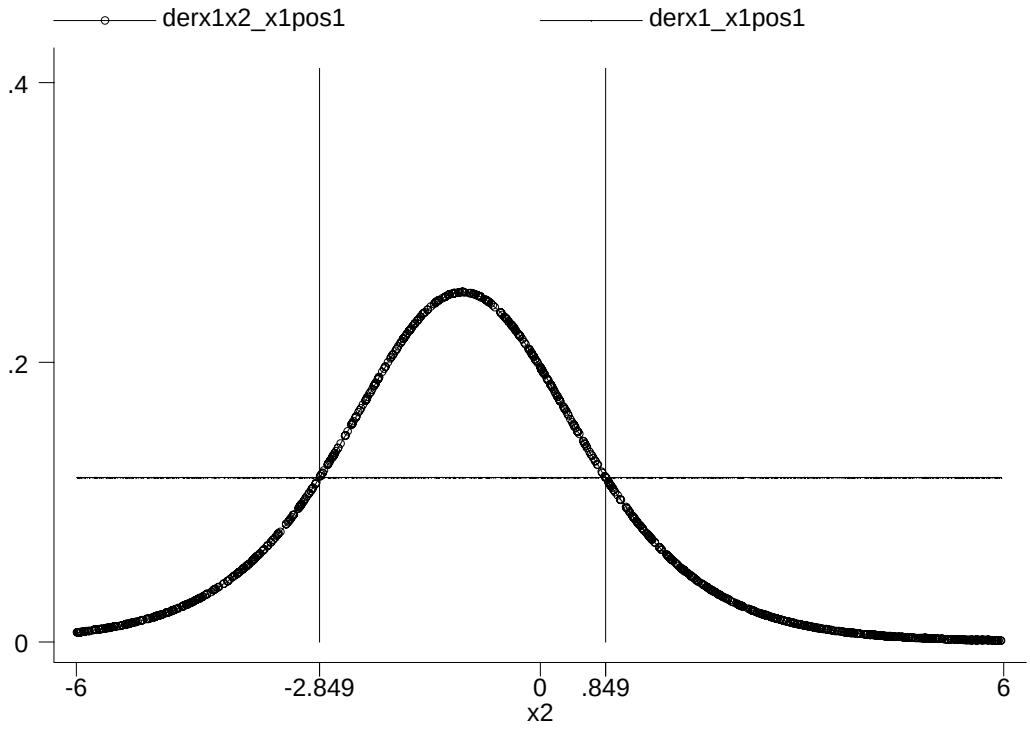

Figure 21. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = 1$

Vertical lines at values of x_2 `-2.84894` and `.84894` indicate that the averaged derivative and specific values of the derivative are equal.

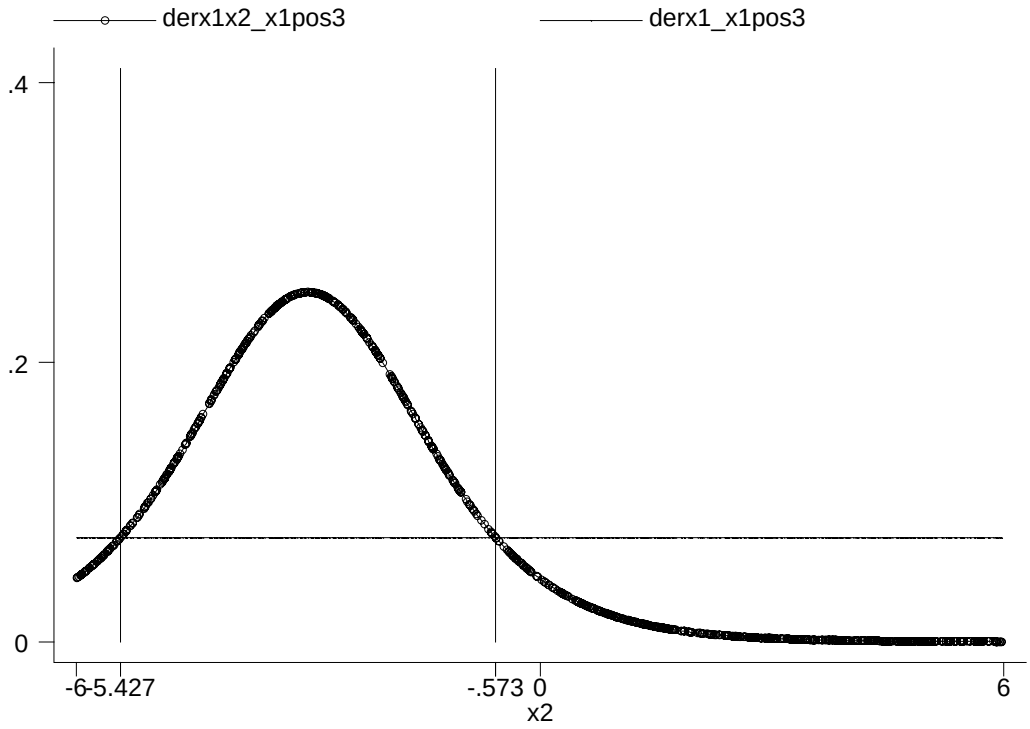

Figure 22. $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1)}{\partial \mathrm{x}_1}$ and $\frac{\partial \Pr ob(d=1 \mid \mathrm{x}_1, \mathrm{x}_2)}{\partial \mathrm{x}_1}$ against x_2 when $\mathrm{x}_1 = 3$

Vertical lines at values of x_2 `-5.42669 and -.57331` indicate that the averaged derivative and specific values of the derivative are equal.

From the graphs one can see that the shortest interval (i.e. the segment of the straight line "inside" of the nonlinear derivative function) is for x_1=0, which is a 'true' minimum. Table 4 displays this a bit more compactly. Here, for each of five values for x_1, the column displays the true conditional derivative after conditioning on both x_1 and x_2 as a function of x_2. The table also highlights the values or x_2 where the conditional on x_2 and unconditional effects coincide.

Table 4

$$\frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1}$$

	$x_1 = -3$	$x_1 = -1$	$x_1 = 0$	$x_1 = 1$	$x_1 = 3$
$x_2 = -6$	.00012	.00091	.00246	.00664	.04517
$x_2 = -5.42669$	.00021	.00161	.00435	.01167	**.07457**
$x_2 = -2.84894$	.00286	.02042	.05174	**.11750**	.24857
$x_2 = -1.76275$	.00839	.05584	**.125**	.21689	.17432
$x_2 = -.84894$	.02042	**.11750**	.20986	.24857	.09336
$x_2 = -.57331$	.02655	.14224	.23053	.23895	**.07457**
$x_2 = 0$	.04517	.19661	.25	.19661	.04517
$x_2 = .57331$	**.07457**	.23895	.23053	.14224	.02655
$x_2 = .84894$	.09336	.24857	.20986	**.11750**	.02042
$x_2 = 1.76275$	.17432	.21689	**.125**	.05584	.00839
$x_2 = 2.84894$	.24857	**.11750**	.05174	.02042	.00286
$x_2 = 5.42669$	**.07457**	.01167	.00435	.00161	.00021
$x_2 = 6$	.04517	.00664	.00246	.00091	.00012
$\int_{-\infty}^{+\infty} \frac{\partial \Pr ob(d=1 \mid x_1, x_2)}{\partial x_1} dG(x_2) = \frac{\partial \Pr ob(d=1 \mid x_1)}{\partial x_1}$	**.07457**	**.11750**	**.125**	**.11750**	**.07457**
Cutoff points (of x_2)	(.57331; 5.42669)	(-.84894; 2.84894)	(-1.76275; 1.76275)	(-2.84894; .84894)	(-5.42669; -.57331)

References

Baker, M and A. Melino, 2000, "Duration dependence and nonparametric heterogeneity: A Monte Carlo study**,"** *Journal of Econometrics*, Vol. 96, No. 2, pp., 357-393.

Cecchetti, S., 1986, "The Frequency of Price Adjustment: A Study of the Newsstand Prices of Magazines," *Journal of Econometrics*, Vol. 31, No. 3, pp. 255-274.

Chamberlain, G., 1983, "Panel Data," in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Amsterdam: North Holland.

Fenton, V., and A. R. Gallant, 1996, "Qualitative and Asymptotic Performance of SNP Density Estimators," *Journal of Econometrics*, Vol. 74, 77–118.

Goldstein, H., and J. Rasbash, 1996, "Improved Approximations for Multilevel Models with Binary Responses," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 159, No. 3., pp. 505-513.

Greene, W., 2000, *Econometric Analysis*, Fourth Edition, Upper Saddle Rive, N.J.: Prentice Hall.

Hausman, J., 1978, "Specification Tests in Econometrics," *Econometrica*, Vol. 46, pp. 1251-1271.

Mroz, T., 1999, "Discrete Factor Approximation in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome," *Journal of Econometrics*, Vol. 92, pp. 233–274.

Neuhaus, J.M., J.D. Kalbfleisch, and W.W. Hauck (1991) "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, Vol. 59, No. 1, pp.25-35.

Rodríguez, Germán and Goldman, Noreen (1995). "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 158, No. 1., pp. 73-89.

Rodríguez, Germán and Goldman, Noreen (2001). "Improved estimation procedures for multilevel models with binary response: a case-study," *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 164, No. 2., pp. 339-355.