IAW-Diskussionspapiere

IAW-Discussion Paper

10

Randomized Response and the Binary Probit Model

Gerd Ronning

August 2003

INSTITUT FÜR
ANGEWANDTE
WIRTSCHAFTSFORSCHUNG
Ob dem Himmelreich 1
72074 Tübingen

T: (0 70 71) 98 96-0
F: (0 70 71) 98 86-99
E-Mail: iaw@iaw.edu
Internet: www.iaw.edu

ISSN: 1617-5654

IAW-Diskussionspapiere

Das Institut für Angewandte Wirtschaftsforschung (IAW) Tübingen ist ein unabhängiges außeruniversitäres Forschungsinstitut, das am 17. Juli 1957 auf Initiative von Professor Dr. Hans Peter gegründet wurde. Es hat die Aufgabe, Forschungsergebnisse aus dem Gebiet der Wirtschafts- und Sozialwissenschaften auf Fragen der Wirtschaft anzuwenden. Die Tätigkeit des Instituts konzentriert sich auf empirische Wirtschaftsforschung und Politikberatung.

Dieses **IAW-Diskussionspapier** können Sie auch von unserer IAW-Homepage als pdf-Datei herunterladen:

http://www.iaw.edu/Publikationen/IAW-Diskussionspapiere

ISSN 1617-5654

Weitere Publikationen des IAW:

- IAW-News (erscheinen 3-4x jährlich)
- IAW-Report (erscheinen 2x jährlich)
- IAW-Wohnungsmonitor Baden-Württemberg (erscheinen 4x jährlich kostenlos)
- IAW-Forschungsberichte

Möchten Sie regelmäßig eine unserer Publikationen erhalten, dann wenden Sie sich bitte an uns:

IAW Tübingen, Ob dem Himmelreich 1, 72074 Tübingen,
Telefon 07071 / 98 96-0
Fax 07071 / 98 96-99
E-Mail: iaw@iaw.edu

Aktuelle Informationen finden Sie auch im Internet unter: http://www.iaw.edu

---

Randomized response and the binary probit model

Gerd Ronning
Department of Economics, University of Tübingen*

18 August 2003

Key Words: Asymptotic Efficiency; Maximum Likelihood; Post Randomisation; Statistical Disclosure.
JEL classification: C21, C25, C42, C81

Summary

The paper analyzes effects of randomized response with respect to some binary dependent variable on the estimation of the probit model. This approach is used in interviews when asking sensitive questions. Alternatively randomization can be considered as a means of statistical disclosure control which has been termed post randomization method (PRAM). The paper shows that all properties concerning parameter estimation are maintained although there is a loss in (asymptotic) efficiency.

*Address: Gerd Ronning, Department of Economics, University of Tübingen, Mohlstr. 36, 72074 Tübingen, phone: ++49 7071 2972571, fax: ++49 7071 295546, email: gerd.ronning@uni-tuebingen.de.

1. Introduction

The binary probit model (see, for example, Greene 2003 chapter 21.3 or Ronning 1991 chapter 2.2.1) considers the effect of some explanatory variable x on a latent continuous variable Y^*, i.e. we assume that the model

$$Y^* \quad = \quad \alpha \, + \, \beta \, x \, + \, \varepsilon \tag{1}$$

holds where ε is a normally distributed random error with $E(\varepsilon) \, = \, 0$ and $V(\varepsilon) \, = \, 1$. However Y^* is observed only as a binary or dichotomous variable Y which is defined by the threshold model

$$Y \quad = \quad \begin{cases} 1 & \text{if} \quad Y^* \, > \, 0 \\ 0 & \text{if} \quad Y^* \, \leq \, 0 \end{cases} \quad . \tag{2}$$

The sample information is given by n pairs $(x_i \, , \, y_i)$ where $y_i \quad \in \quad \{0, 1\}$ and x_i is an arbitrary real number. Maximum likelihood estimation of the two unknown parameters α and β is straightforward; see some standard text as, for example, Greene (2003) or Ronning (1991). Note that we have already introduced the usual identifying restrictions, i.e. zero threshold and unit error variance. We confine ourselves to the case of just one regressor which is assumed to be continuous. The results in this paper however apply also to the more general case of an arbitrary number of explanatory variables after minor modifications. Here we consider randomization of the dichotomous variable y which switches its values with some prescribed transition probability (leaving the explanatory variable x in its original form). In the following section the method is described in some more detail. Section 3 then considers the effect on the estimation of the binary probit model. Some concluding remarks are added in section 4.

2. Randomized response and post randomization

Randomized response originally was introduced to avoid non-response in surveys containing sensitive questions on, e.g., drug consumption or AIDS disease. See Warner (1965). Särndal et. al. (1992 p. 573) suggested use of this method "to protect the anonymity of individuals". A good description of the difference between the two (formally equivalent) approaches is given by van den Hout and van der Heijden (2002): In the randomized response setting the stochastic model has to be defined in advance of data collection whereas in post randomization this method will be applied to the data already obtained.

For the case of a dichotomous variable the method can be described as follows: Consider a $(2 \, \times \, 2)$ probability matrix describing transition probabilities for the two states "0" and "1" of a random variable Y:

$$P_y \quad = \quad \begin{pmatrix} p_{00} & p_{01} \\ \\ p_{10} & p_{11} \end{pmatrix}$$

In the following we denote the dichotomous variable obtained from post randomization by Y^m which we call the 'masked' variable. Then the transition probabilities can

be defined by $p_{jk} \equiv P(Y^m = j \,|\, Y = k)$ with $j, k \,\varepsilon\, \{0,1\}$ and $p_{j0} + p_{j1} = 1$ for $j = 0, 1$. Since there is no argument not to treat the two states symmetrically, we define by π the probability of no change. Then the probability matrix can be written as follows:

$$P_y = \begin{pmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{pmatrix}$$

Note that this matrix is singular if $\pi = \frac{1}{2}$ which will become important later on.

When the sample of the dependent variable has undergone randomization, we will have n observations y_i^m where y_i^m is the dichotomous variable obtained from y_i by the randomization procedure.

Randomization has the advantage that the original distribution of Y can be estimated from the masked observations y_i^m. See Kooiman et al (1997) for a detailed exposition. The sample of unmasked observations is completely characterized by n, the number of observations, and $\theta = \sum_i y_i$, the number of 'successes' which is the parameter of interest. Defining $T^m = \sum_i Y_i^m$ an unbiased estimator is given by

$$\hat{\theta} = \frac{T^m - n\,(1-\pi)}{(2\pi - 1)}$$

with variance

$$var(\hat{\theta}) = \frac{n\,\pi\,(1-\pi)}{(2\pi-1)^2} \tag{3}$$

which does not depend on θ. However the coefficient of variation is inversely related to θ and therefore distortion from post-randomization will be serious if this parameter is near to 0 or n. "This fits nicely into our general purpose to protect rare scores, since these are most vulnerable to disclosure." (Kooiman et al 1997 p. 4).

3. Estimation of the probit model under randomization

Let us now turn to the estimation of the probit model as given in (1) and (2). From figure 1 it is apparent that under randomization we have the following data generating process:

$$Y_i^m = \begin{cases} 1 & \text{with probability} \quad \Phi_i\,\pi + (1-\Phi_i)\,(1-\pi) \\ 0 & \text{with probability} \quad \Phi_i\,(1-\pi) + (1-\Phi_i)\,\pi \end{cases} \tag{4}$$

Here Φ_i denotes the conditional probability under the normal distribution that the unmasked dependent variable Y_i takes on the value 1 for given x_i, i.e. $\Phi_i \equiv \Phi(\alpha + \beta x_i) = P(Y_i^* > 0 \,|\, x_i)$. See (1) and (2).

From (4) we obtain the following likelihood function:

$$\mathcal{L}(\alpha, \beta | (y_i^m, x_i), i = 1, \ldots, n)$$
$$= \prod_{i=1}^{n} \left(\Phi_i\,\pi + (1-\Phi_i)\,(1-\pi)\right)^{y_i^m} \left(\Phi_i\,(1-\pi) + ((1-\Phi_i)\,\pi\right)^{(1-y_i^m)} \quad .$$

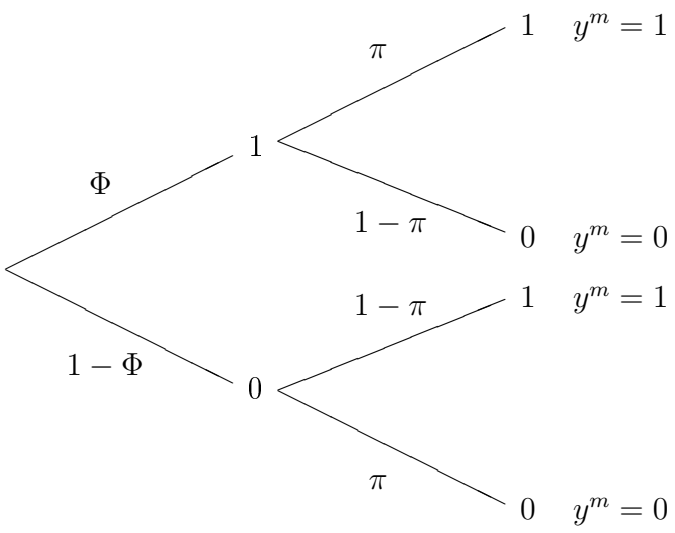

Figure 1: Data generating process under randomization

Global concavity of this function with respect to α and β may be checked by deriving first and second (partial) derivatives of the log-likelihood function

$$L \equiv \log(\mathcal{L}) = \sum_i^n y_i^m \log\left[\Phi_i \pi + (1-\Phi_i)(1-\pi)\right] + (1 - y_i^m) \log([\Phi_i (1-\pi) + ((1-\Phi_i)\pi] .$$

The partial first-order derivatives with respect to α and β are given by

$$\begin{aligned}
\frac{\partial L}{\partial \alpha} &= (2\pi - 1) \sum_i \left[y_i^m \frac{\phi_i}{W_i} - (1 - y_i^m) \frac{\phi_i}{1 - W_i} \right] \\
&= (2\pi - 1) \sum_i \left[\frac{(y_i^m - W_i)\,\phi_i}{W_i\,(1 - W_i)} \right] \\
\frac{\partial L}{\partial \beta} &= (2\pi - 1) \sum_i \left[y_i^m \frac{\phi_i}{W_i} - (1 - y_i^m) \frac{\phi_i}{1 - W_i} \right] x_i \\
&= (2\pi - 1) \sum_i \left[\frac{(y_i^m - W_i)\,\phi_i}{W_i\,(1 - W_i)} \right] x_i \qquad (5)
\end{aligned}$$

where we use the following definitions:

$$\phi_i \equiv \phi(\alpha + \beta x_i) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} (\alpha + \beta x_i)^2 \right\} \quad (\text{ standard normal density})$$

and

$$W_i \equiv \pi \Phi_i + (1 - \pi)(1 - \Phi_i) \qquad (\text{probability of observing } y_i^m = 1) .$$

Note that the first order conditions would be fulfilled for $\pi = \frac{1}{2}$ for *any* α and β leaving these two parameters unidentified. Therefore we exclude this case in the following by assuming

$$\pi \quad \neq \quad \frac{1}{2} \quad . \tag{6}$$

Moreover the two partial derivatives are similar (with W_i instead of Φ_i) in structure to those concerning the standard probit model disregarding the proportional factor $(2\pi - 1)$. See, e.g. , Ronning (1991 p. 45).

Unfortunately it turns out that the Hessian matrix formed from the second order partial derivatives is no longer negative definite contrary to the standard probit case which guarantees global concavity of the log-likelihood function. First note that

$$\begin{array}{lcllcl}
\frac{\partial W_i}{\partial \alpha} & = & (2\pi - 1)\,\phi_i & \frac{\partial W_i}{\partial \beta} & = & (2\pi - 1)\,\phi_i\,x_i \\
\frac{\partial (1-W_i)}{\partial \alpha} & = & -(2\pi - 1)\,\phi_i & \frac{\partial (1-W_i)}{\partial \beta} & = & -(2\pi - 1)\,\phi_i\,x_i \\
\frac{\partial W_i(1-W_i)}{\partial \alpha} & = & (2\pi - 1)\,\phi_i\,(1 - 2W_i) & \frac{\partial W_i(1-W_i)}{\partial \beta} & = & (2\pi - 1)\,\phi_i\,x_i\,(1 - 2W_i)
\end{array}$$

and

$$\begin{array}{rcccl}
\frac{\partial\,\phi_i}{\partial \alpha} & = & -(\alpha + \beta x_i)\,\phi_i & = & -z_i\,\phi_i \\
\frac{\partial\,\phi_i}{\partial \beta} & = & -(\alpha + \beta x_i)\,\phi_i\,x_i & = & -z_i\,\phi_i\,x_i
\end{array}$$

using $z_i \equiv \alpha + \beta x_i$.

For the second-order partial derivative with respect to α we obtain

$$\begin{aligned}
\frac{\partial^2 L}{(\partial\alpha)^2} &= (2\pi-1)\left\{\sum_i\left[\left(\frac{\partial}{\partial\alpha}\,\frac{y_i^m - W_i}{W_i(1-W_i)}\right)\phi_i + \frac{y_i^m - W_i}{W_i(1-W_i)}\left(\frac{\partial}{\partial\alpha}\,\phi_i\right)\right]\right\} \\
&= (2\pi-1)\left\{\sum_i\left[\frac{\phi_i(2\pi-1)(-W_i^2 - y_i^m(1-2W_i)) - W_i(1-W_i)(y_i^m - W_i)z_i}{(W_i(1-W_i))^2}\right]\phi_i\right\} \\
&= -(2\pi-1)\left\{\sum_i\left[\frac{(2\pi-1)(y_i^m - 2y_i^m W_i + W_i^2)\,\phi_i + (y_i^m - W_i)W_i(1-W_i)z_i}{(W_i(1-W_i))^2}\right]\phi_i\right\} \\
&= -(2\pi-1)\left\{\sum_i \frac{g_i\,\phi_i}{(W_i(1-W_i))^2}\right\} \qquad (7)
\end{aligned}$$

where

$$g_i \quad = \quad (2\pi-1)(y_i^m - 2y_i^m W_i + W_i^2)\,\phi_i + (y_i^m - W_i)W_i(1-W_i)z_i \quad . \tag{8}$$

Corresponding results are obtained for the two other second-order partial derivatives leading to the following Hessian matrix:

$$H^m \quad = \quad -(2\pi-1)\sum_i \frac{g_i\,\phi_i}{(W_i(1-W_i))^2}\,\mathbf{u}_i\,\mathbf{u}_i' \tag{9}$$

with $\mathbf{u}_i' = \begin{pmatrix} 1 & x_i \end{pmatrix}$. Since $2\pi - 1$ may be either positive or negative and the function g_i in (8) is more complex than in the standard case, the proof of negative definiteness used in the standard probit case does not go through here. See, for example, Amemiya(1985 p. 274) or Ronning (1991 p. 46).

However we obtain a simple formula for the information matrix from which it is immediately apparent that estimation under randomization implies an efficiency loss. First note that $\mathrm{E}\,[Y_i^m] = W_i$ and therefore the expected value of g_i as a function of Y_i^m is given by

$$\begin{aligned}
\mathrm{E}\,[g(Y_i^m)] &= \mathrm{E}\left[(2\pi-1)(Y_i^m - 2Y_i^m W_i + W_i^2)\,\phi_i + (Y_i^m - W_i)W_i(1-W_i)z_i\right] \\
&= (2\pi-1)\,(W_i - 2W_i^2 + W_i^2)\phi_i \;+\; (W_i - W_i)W_i(1-W_i)z_i \\
&= (2\pi-1)\,W_i\,(1-W_i)\,\phi_i
\end{aligned}$$

Therefore the information matrix (or the the expected value of the Hessian matrix multiplied by -1) in case of masked data is given by

$$\mathcal{I}^m \;=\; (2\pi-1)^2 \sum_i \frac{\phi_i^2}{W_i\,(1-W_i)}\,\mathbf{u}_i\,\mathbf{u}_i' \tag{10}$$

whereas in the case of unmasked data we obtain (e.g. Amemiya 1985 p. 272 or Ronning 1991 p. 46)

$$\mathcal{I} \;=\; \sum_i \frac{\phi_i^2}{\Phi_i\,(1-\Phi_i)}\,\mathbf{u}_i\,\mathbf{u}_i' \tag{11}$$

We now want to show that the difference $\mathcal{I} - \mathcal{I}^m$ is nonnegative definite. It is sufficient to show that for every i

$$\frac{1}{\Phi_i\,(1-\Phi_i)} \;>\; \frac{(2\pi-1)^2}{W_i\,(1-W_i)} \tag{12}$$

or $W_i\,(1-W_i) \;>\; (2\pi-1)^2\,\Phi_i\,(1-\Phi_i)$ which follows immediately from

$$W_i\,(1-W_i) \;=\; (2\pi-1)^2\,\Phi_i\,(1-\Phi_i) \;+\; \pi(1-\pi)$$

recalling the definition $W_i = \pi\Phi_i + (1-\pi)(1-\Phi_i)$ from above.

Which values of π imply the largest efficiency loss ? For the 'weights' in (10) we can write

$$\begin{aligned}
\frac{(2\pi-1)^2}{W_i\,(1-W_i)} &= \frac{(2\pi-1)^2}{(2\pi-1)^2\,\Phi_i\,(1-\Phi_i) \;+\; \pi(1-\pi)} \\
&= \frac{1}{\Phi_i(1-\Phi_i) \;+\; \frac{\pi(1-\pi)}{(2\pi-1)^2}} \quad .
\end{aligned}$$

Since the function

$$h(\pi) \;=\; \frac{\pi(1-\pi)}{(2\pi-1)^2}$$

is symmetric, i.e. $h(\pi) = h(1-\pi)$, monotonically increasing and tending towards infinity at $\pi = 1/2$, the weights in (10) tend towards zero and the information matrix $\mathcal{I}^m$ tends towards a zero matrix when π (or $(1-\pi)$ tends towards $1/2$. Note that the function $h(\pi)$ has already appeared in the variance of $\hat{\theta}$ in section 2.

4. Concluding remarks

The paper has shown that randomization of the binary dependent variable involves an efficiency loss which is larger when the probability of switching tends towards 1/2 whereas this loss is small when this probability is near zero or one. We have not discussed performance of the maximum likelihood estimator since it becomes clear from the presented results that consistency and asymptotic normality still hold under randomization of the dependent variable. From the perspective of statistical disclosure control masking of the explanatory variable should also be taken into account. For example, the continuous explanatory variable could be masked by microaggregation or addition of noise. See, for example, Domingo-Ferrer (2002), in particular the section on microdata protection. However, this additional transformation seems to have no clear-cut effects on estimation but will be a topic of further research.

Acknowledgements

Research in this paper is related to the project "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" financed by German Ministry of Research and Technology. I have to thank Nathalie Höllig for preparing the figure.

References

Amemiya, T. , 1985 , Advanced Econometrics (Basil Blackwell: Oxford).

Domingo-Ferrer , J. (editor), 2002, Inference Control in Statistical Databases (Springer, Berlin).

Greene, W.H. , 2003 , Econometric Analysis (Prentice Hall: Upper Saddle River) , fifth edition.

Kooiman, P., L. Willenborg and J. Gouweleeuw, 1997, PRAM: a method for disclosure limitation of micro data. http://www.cbs.nl/sdc/ruis.htm

Ronning, G. ,1991 , Mikroökonometrie (Springer, Berlin).

Särndal, C.-E., B. Swensson and J. Wretman , 1992 , Model Assisted Survey Sampling (Springer , New York).

van den Hout, A. and P.G.M. van der Heijden, 2002 , Randomized response, statistical disclosure control and misclassification: a review, International Statistical Review 70, 2-69-288.

Warner, S.L. , 1965 , Randomized response: a survey technique for eliminating evasive answer bias, Journal of the American Statistical Association 57, 622-627.

IAW-Diskussionspapiere

Bisher erschienen:

Nr. 1
Das Einstiegsgeld – eine zielgruppenorientierte negative Einkommensteuer: Konzeption, Umsetzung und eine erste Zwischenbilanz nach 15 Monaten in Baden-Württemberg
Sabine Dann / Andrea Kirchmann / Alexander Spermann / Jürgen Volkert

Nr. 3
Gut betreut in den Arbeitsmarkt? Eine mikroökonomische Evaluation der Mannheimer Arbeitsvermittlungsagentur
Jürgen Jerger / Christian Pohnke / Alexander Spermann

Nr. 4
Das IAW-Einkommenspanel und das Mikrosimulationsmodell SIMST
Peter Gottfried / Hannes Schellhorn

Nr. 5
A Microeconometric Characterisation of Household Consumption Using Quantile Regression
Niels Schulze / Gerd Ronning

Nr. 6
Determinanten des Überlebens von Neugründungen in der baden-württembergischen Industrie – eine empirische Survivalanalyse mit amtlichen Betriebsdaten
Harald Strotmann

Nr. 7
Die Baulandausweisungsumlage als ökonomisches Steuerungsinstrument einer nachhaltigkeitsorientierten Flächenpolitik
Raimund Krumm

Nr. 8
Making Work Pay: U.S. American Models for a German Context?
Laura Chadwick, Jürgen Volkert

Nr. 9
Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik
Martin Rosemann

Nr. 10
Randomized Response and the Binary Probit Model
Gerd Ronning