Available online at www.sciencedirect.com

SCIENCE DIRECT®

Economics Letters 86 (2005) 221–228

economics letters

www.elsevier.com/locate/econbase

Randomized response and the binary probit model

Gerd Ronning*

Department of Economics, University of Tübingen, Mohlstr. 36, 72074 Tübingen, Germany

Received 26 September 2003; received in revised form 12 July 2004; accepted 15 July 2004
Available online 10 November 2004

Abstract

The paper analyzes effects of randomized response with respect to some binary dependent variable on the estimation of the probit model. Alternatively, randomization can be considered as a means of statistical disclosure control which has been termed post randomization method (PRAM).

Keywords: Asymptotic efficiency; Maximum likelihood; Misclassification; Post randomization; Statistical disclosure

JEL classification: C21; C25; C42; C81

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

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

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

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

* Tel.: +49 7071 2972571; fax: +49 7071 295546.
E-mail address: gerd.ronning@uni-tuebingen.de.

0165-1765/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.econlet.2004.07.013

The sample information is given by n pairs (x_i,y_i) where $y_i \in \{0,1\}$ 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. In Section 4, we describe briefly the close relation of randomization to the literature on estimation of the probit model under misclassification. Some concluding remarks are added in Section 5.

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.

Randomization of the binary variable Y can be described as follows: Let Y^m denote the ‘masked’ variable obtained from post randomization. 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 $j=0,1$. If we define the two probabilities of no change by $p_{00} \equiv \pi_0$ and $p_{11} \equiv \pi_1$, respectively, the probability matrix can be written as follows:

$$P_y = \begin{pmatrix} \pi_0 & 1-\pi_0 \\ 1-\pi_1 & \pi_1 \end{pmatrix}$$

Note that this matrix is singular if $\pi_0+\pi_1=1$ which will become important later on. Since in the post randomization procedure the two probabilities are known and there is no argument not to treat the two states symmetrically, in the following we will consider the special case

$$\pi_0 = \pi_1 \tag{3}$$

but will come back to the more general case later on in Section 4.

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

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