X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

A COMPARISON OF SEMIPARAMETRIC ESTIMATORS FOR THE BINARY CHOICE MODEL

Alicia Pérez-Alonso[1], Maja Rynko[1]

[1]Research Group in Economic Analysis and Departamento de Estatística e Investigación Operativa, Universidade de Vigo

[2]Polish Central Statistical Office and Department of Economics, European University Institute

ABSTRACT

This paper compares five semi-parametric estimators for binary outcome models: a 'semi-nonparametric' estimator, a maximum score estimator, a smooth version of the maximum score estimator, a kernel-based quasi maximum likelihood estimator and a least squares estimator for the transformed dependent variable. Monte Carlo evidence comparing the performance of the estimators is presented. The empirical example models the propensity to smoke and illustrates the differences in parameters estimates across analysed frameworks.

Keywords: Binary choice; semiparametric estimation; identification; single-index; latent variable; smoking

1. INTRODUCTION

In this paper, we compare five different semiparametric estimators of binary response model. We classify the estimators into three different classes depending on their identification conditions: estimators based on the assumption of *quantile independence*; estimators based on the assumption of *index sufficiency*; and estimators based on the assumption of *partial independence and uncorrelated errors*. In the first class, we consider the maximum score estimator of Manski (1975, 1985), and its smoothed version developed by Horowitz (1992). The second class includes the seminonparametric estimator (SNP) of Gallant and Nychka (1987) adapted for the binary response case by Gabler et al. (1993), and the quasi-maximum-likelihood semiparametric estimator of Klein and Spady (1993). We study also the estimator of Lewbel (2000), which is assigned to the third class.

The five estimators under study have received a lot of attention in the literature on semiparametric estimation. Using the work trip mode choice model, Horowitz (1993) compares the probit model with the estimators of Manski, the smooth maximum score and Klein and Spady. Gerfin (1996) studies the decision of labor force participation of married women using the probit model, the smoothed maximum score, the SNP and Klein and Spady estimators. Various authors have also adopted selected semiparametric estimators to the case of ordinal dependent variable. Stewart (2005) assesses the ordered probit model, the estimator of Lewbel and the SNP estimator, among others. Melenberg and van Soest (1996) analyse the SNP estimator and a generalization of the smooth maximum score estimator.

Parametric methods such as Probit and Logit are still the most commonly used methods in empirical work even if their assumptions are often difficult to justify. Their popularity may be explained to a big extent by their availability in every statistical software, while the semiparametric or fully nonparametric specifications are often neglected. For instance, the Stata software provides only the procedures for the SNP and Klein and Spady estimators, whereas R software - the *np package* of Hayfield and Racine (2008) contains the estimator of Klein and Spady and an alternative approach of Ichimira (1993). To our knowledge, the Manski's maximum score estimator is included only in econometric software Limdep, that contains also Klein and Spady semiparametric approach. Thus, we also provide an unified R code for the five estimators that we discuss in this paper, which can be obtained from the authors upon request.

2. THE MODEL AND IDENTIFICATION

We consider the binary choice model of the general form:

$$Y = \mathbf{1}(\nu + X'\beta + e \geq 0), \qquad (1)$$

where Y is a binary observable variable, ν is a continuous regressor whose coefficient is set equal to one by convention, X is a k-dimensional vector of explanatory variables, $\beta \in \mathbb{R}^k$ is the vector of coefficients of interests, e is a scalar error term, and $\mathbf{1}(\cdot)$ is the indicator function that equals one if its argument is true and zero otherwise. Hereafter, $F(\xi)$ and $Q_\alpha(\xi)$ denote the cumulative distribution function (cdf) and the α-th quantile of ξ, respectively.

Manski (1988) derives primitive conditions that ensure identifiability of β in the semiparametric binary choice model. The kind of conditions that are required crucially depends on the additional assumptions about the error term e that are made in analysed estimation strategies. Basically, three semiparametric settings are analyzed:

i) *Statistical Independence*: $F(e|\nu, X) = F(e)$ and $Q_{0.5}(e) = 0$;
ii) *Index Sufficiency*: $F(e|\nu, X) = F(e|\nu + X'\beta)$;
iii) *Quantile Independence*: $Q_\alpha(e \mid \nu, X) = Q_\alpha(e)$ for a given α.

Note that when the index sufficiency assumption is met then (1) is a single-index model, since $E[Y = 1 \mid \nu, X]$ depends on (ν, X) only through $\nu + X'\beta$. Lewbel (2000) provides a fourth setting:

iv) *Partial Independence + Uncorrelated Error*: $F(e|\nu, X) = F(e|X)$ and $E[X'e] = 0$.

As pointed out in Magnac and Maurin (2007), the partial independence assumption overcomes Manski's fundamental imposibility result according to which *mean independence* (i.e. $E[e|\nu, X] = 0$) is not sufficient for identifying β no matter what conditions on the support of (ν, X) are adopted (see Manski 1988). In all settings, it is required that the special regressor ν has a large support, though this condition is relaxed by Magnac and Maurin (2007) for the fourth setting.

The main conclusions of the analysis in Manski (1988) are:

a) in all settings, parameters can only be identified up to scale;
b) in the index sufficiency setting, an intercept term is not identified, but in the other two settings the intercept term can be identified up to scale[1];
c) in the index sufficiency setting, parameters can only be identified up to sign and, in addition, smoothness conditions on the real function $h(z) \equiv F_z(z)$ are required, where for $z \in \mathbb{R}$, F_z denotes the cdf of $e \mid \nu + X'\beta = z$;
d) in the statistical independence setting, the cdf of e is identified.

Note that the first setting is more restrictive than the third one, and it is also more restrictive than the second one plus a location assumption. Moreover, while *quantile independence* assumes that one quantile of e is independent of all covariates, partial independence is equivalent to assuming that all quantiles of e are independent of the

[1]The index sufficiency setting also allows for identification of the intercept term if location assumptions about the cdf of e are made (see Chen 1999).

special regressor ν, so both settings seem comparably restrictive in this sense.

If we compare these assumptions with those that are usually made in a parametric framework, we observe that the most important additional assumption that is required in the semiparametric framework is the existence of a significant continuous regressor. Note that in parametric models the intercept term is identified since the mean of e is assumed to be zero, and the scale is fixed since the variance of e is assumed to be known[2]. In addition, parametric models assume that the errors and regressors are independent and thus do not allow for heteroskedasticity. The index sufficiency setting only allows for very specific types of heteroskedasticity, where the regressors affect the conditional variance $Var(e|\nu, X)$ through the single-index $\nu + X'\beta$. The quantile independence setting permits much more general forms of conditional heteroskedasticity.

3. CONCLUSIÓNS

In the empirical example, the estimated coefficients give generally a consistent picture about the sign, denoting the direction of influence between the socio-demographic characteristics and the individual's propensity to not-smoke. The gender plays a significant role with males being more likely to smoke. Higher education, greater income, being married and being religious decrease the probability of smoking.

However, the magnitudes of these effects differ across the models. Probit, Manski, Horowitz and SNP specifications estimate the effect of one additional year of education as substantially weaker than being religious, whereas in Klein-Spady specification the difference in magnitudes between those two variables is much smaller. The results concerning gender and marital status effects differ across the models. In probit specification, being married offsets the negative effect of gender, which is not true for the semiparametric models. The striking difference relates to the Klein-Spady estimator, for which the marital status has much higher influence in absolute terms on propensity to smoke than the gender, whereas the opposite is observed for Manski, Horowitz and SNP.

The presented results do not provide the answer which specification constitutes the right approach. This decision is unfortunately left to the subjective choice of the researcher.

[2] The identification restrictions imposed directly on the coefficients applied in the semiparametric framework could also be introduced in the parametric probit specification: the exclusion of the constant term (instead of imposing the error term to have mean 0) or the requirement of one coefficient to equal 1 (instead of restricting the variance to be equal to 1)

REFERENCES

Chen, S. (1999), Semiparametric Estimation of a Location Parameter in the Binary Choice Model, *Econometric Theory* 15, 79-98.

Gabler, S., Laisney, F. and Lechner, M. (1993), Seminonparametric estimation of binary-choice models with an application to labor-force participation, *Journal of Business & Economic Statistics* 11(1), 61-80.

Gallant R. A., Nychka D.W. (1987), Semi-Nonparametric Maximum Likelihood Estimation, *Econometrica* 55(2), 363-390.

Gerfin, M. (1996), Parametric and semiparametric estimation of the binary response model of labour market participation, *Journal of Applied Econometrics* 11, 321-339.

Hayfield, T. and Racine, J., Nonparametric Econometrics: The np Package, *Journal of Statistical Software*, 27(5), 1-32.

Horowitz, J.L. (1992), A smoothed maximum score estimator for the binary response model, *Econometrica* 60(3), 505-531.

Horowitz, J.L., (1993), Semiparametric estimation of a work-trip choice model, *Journal of Econometrics* 58, 49-70.

Ichimura, H. (1993), Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics* 58, 71-120.

Klein, R.W. and Spady, R.H. (1993), An Efficient Semiparametric Estimator for Binary Response Models, *Econometrica* 61(2), 387-421.

Lewbel, A. (2000), Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables, *Journal of Econometrics* 97, 145-177.

Magnac, T., and Maurin, E. (2007), Identification and information in monotone binary models, *Journal of Econometrics* 139, 76-104.

Manski, C.F. (1975), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics* 3, 205-228.

Manski, C.F. (1985), Identification of Binary Response Models, *Journal of the American Statistical Association* 83, 729-738.

Manski, C.F. (1988), Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics* 3, 205-228.

Melenberg, B. and van Soest, A. (1996), Measuring the cost of Children: Parametric and Semiparametric Estimators, *Statistica Neerlandica*, 50, 171-192.

Stewart, M.B. (2005), A comparison of semiparametric estimators for the ordered response model, *Computational Statistics & Data Analysis* 49, 555-573.