DIW Diskussionspapiere
Discussion Papers

Discussion Paper No. 215

Evaluation of a pseudo-R^2 measure for panel probit models

by
Martin Spiess

Berlin, May 2000

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 14195 Berlin
Phone: +49-30-89789- 0
Fax: +49-30-89789- 200
Internet: http://www.diw.de
ISSN 1433-0210

Evaluation of a pseudo-R^2 measure for panel probit models

Martin Spiess

DIW, GSOEP

Abstract

A simulation study designed to evaluate the pseudo-R^2_T proposed by Spiess and Keller (1999) suggests that this measure represents the goodness-of-fit not only of the systematic part, but also of the assumed correlation structure in binary panel probit models.

Key words: Goodness-of-fit; Pseudo-R^2; Panel probit model; Simulation study
JEL classification: C23, C25

1 Introduction

The coefficient of determination, usually denoted as R^2, is used to assess the goodness-of-fit of univariate linear regression models in many applications. Assuming a univariate latent linear model, a generalisation of R^2 to univariate probit models with ordered categorical responses is given by McKelvey and Zavoina (1975). In several simulation studies comparing different goodness-of-fit measures, this so-called pseudo-R^2 was closest to R^2 of the underlying linear model (e.g. Hagle and Mitchell II, 1992; Veall and Zimmermann, 1992, 1996; Windmeijer, 1995). A generalisation of R^2 to multivariate linear models with invariant covariates within clusters but different parameter values is given by the squared

*Corresponding author.
I would like to thank Markus Pannenberg for helpful comments.
Martin Spiess, DIW, GSOEP, Koenigin-Luise-Str. 5, 14195 Berlin, Germany, Phone: +49-30-897 89 602, Fax: +49-30-897 89 200, e- mail: mspiess@diw.de

trace correlation denoted as R^2_T (Hooper, 1959). Spiess and Keller (1999) propose a generalization of this measure to general multivariate probit models with ordered categorical responses which can be calculated even if the estimators are calculated using non-maximum likelihood methods. However, they do not discuss or evaluate the properties of this measure, denoted as pseudo-R^2_T.

In this paper, the results of a simulation study are presented to address the following question: Does this measure represent the goodness-of-fit of the estimated model 'appropriately' ? As a special case of the general multivariate probit model, the binary panel probit model is considered.

2 Pseudo-R^2_T

For the binary panel model considered, let N be the number of subjects ($n = 1, \ldots, N$), T be the number of observations for every subject ($t = 1, \ldots, T$) and $\mathbf{y}_n = (y_{n1}, \ldots, y_{nT})'$ denote the $(T \times 1)$ vector of binary responses for the nth subject. Furthermore, let $\mathbf{x}_{nt}$ denote the $(P \times 1)$ vector of covariates associated with the tth observation of the nth subject, and $\mathbf{X}_n$ the $(T \times P)$ matrix of covariates associated with the nth subject. In what follows, a threshold model

$$y^*_{nt} = \mathbf{x}'_{nt}\beta + \epsilon_{nt} \quad \text{and} \quad y_{nt} = \begin{cases} 1 & \text{if} \quad y^*_{nt} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

is assumed, where y^*_{nt} is the latent continuous response variable, ϵ_{nt} is the error term, and β is the unknown regression parameter vector. It is also assumed that $\epsilon_n \sim N(0, \mathbf{R}(\alpha))$ and $\mathrm{E}(\epsilon_n \mathbf{x}'_{nt}) = \mathbf{0}$ for all n, t, where $\epsilon_n = (\epsilon_{n1}, \ldots, \epsilon_{nT})^T$ and $\mathbf{R}(\alpha)$ is a correlation matrix considered as a function of a parameter α. Observations from different subjects are assumed to be independent. Note that there is no loss of generality in restricting attention to $\mathrm{Cov}(\epsilon_n) = \mathbf{R}(\alpha)$ (Spiess and Keller, 1999).

Given estimates $\hat{\beta}$ and $\mathbf{R}(\hat{\alpha})$, an estimate of the residual sum of squares and product (SSP) matrix of the underlying model is $\widehat{\mathbf{SSP}}_R = N\mathbf{R}(\hat{\alpha})$ and of the fitted SSP matrix is $\widehat{\mathbf{SSP}}_F = \sum_n (\hat{\mathbf{y}}^*_n - \bar{\hat{\mathbf{y}}}^*)(\hat{\mathbf{y}}^*_n - \bar{\hat{\mathbf{y}}}^*)'$, where $\hat{\mathbf{y}}^*_n = \mathbf{X}_n\hat{\beta}$ and $\bar{\hat{\mathbf{y}}}^* = N^{-1}\sum_n \hat{\mathbf{y}}^*_n$. Thus, the total SSP matrix can be approximated by $\widehat{\mathbf{SSP}}_T = \widehat{\mathbf{SSP}}_F + \widehat{\mathbf{SSP}}_R$. The measure pseudo-$R^2_T$ is then given by

$$\text{pseudo-}R^2_T = T^{-1}\mathrm{tr}(\widehat{\mathbf{SSP}}_T^{-1}\widehat{\mathbf{SSP}}_F), \tag{2}$$

where tr($\mathbf{A}$) means trace of matrix $\mathbf{A}$. It is worth noting, however, that generally even in the linear panel model the above partitioning is not entirely valid in finite samples. However, in large samples, the above partitioning holds approximately.

3 Simulation study: Description

Data were generated according to model (1) with $P \in \{3, 4\}$, $T \in \{5, 6, 8\}$ and $N \in \{100, 400, 800\}$. Covariates were generated according to a normal (n), poisson (p), gamma (g), uniform (u) or dichotomous (d; with pr$(x = 1) = .3$) distribution, varying freely over all NT observations. A specific combination of covariates, for example one uniformly, one dichotomously and one normally distributed covariate, will be denoted as exv $=$ u, n, d. The errors were generated according to a multivariate normal distribution with expectation zero. As a correlation structure, either an AR(1) structure with parameter $\rho \in \{0.5, 0.8\}$ or, if $T = 6$, a 'free', Toeplitz-like correlation structure, given by the elements below the diagonal of the correlation matrix $\rho = (\rho_{2,1}, \rho_{3,1}, \rho_{3,2}, \ldots, \rho_{T,T-1})$, where $\rho_{2,1} = \rho_{3,2} = \rho_{4,3} = .9$, $\rho_{5,4} = \rho_{3,1} = \rho_{4,2} = .8$, $\rho_{5,3} = \rho_{6,5} = .7$, $\rho_{4,1} = \rho_{5,2} = \rho_{6,4} = .6$, $\rho_{6,3} = .5$, $\rho_{5,1} = .4$, $\rho_{6,2} = .3$ and $\rho_{6,1} = .1$ was simulated. The goodness-of-fit was varied by multiplying the correlation matrix by different values. To control for approximately similar proportions of $y_t = 1$ for all T (approximately .1, .3, .5 and .9, denoted as $\mathrm{p}_t = .1, .3, .5$ or .9) given the different values of the multiplier, the parameter weighting the constant term with value one generally had to be adjusted accordingly. Parameters were estimated using a GEE-type approach, which allows consistent estimation of regression and correlation structure parameters of the underlying model, the former being consistent even if the correlation structure is misspecified (Spiess, 1998; Spiess and Keller, 1999). Estimators were calculated assuming independence (Ind), an equi- (Equi), an AR(1)- and a Toeplitz correlation structure, respectively. According to each simulated and estimated model, 500 data sets were generated. To compare the values of pseudo-R^2_T with 'true' values, R^2_{0T} was calculated using (2), where instead of estimates, 'true' parameter values and the simulated error covariance matrix were used.

4 Simulation study: Results

Results of the simulation study are given in Figure 1 for different simulated models, defined by the number of subjects (N), the correlation structure (AR(1) and a 'free' correlation structure), the number of observations within subjects (T) and the proportion of $y_t = 1$ (p_t). Generally, no convergence problems occurred. However, if $N = 100$ and $R^2_{0T} \approx 0.85$, estimates converged for less than 475 data sets if an equicorrelation or an AR(1) structure was estimated. Therefore, these results are omitted.

Insert Figure 1 about here

The results given in Figure 1 are in accordance with the results not reported and can be summarized as follows: First, the better the 'fit' of the underlying model, i.e. the smaller the error variance, the higher the value of pseudo-R^2_T. Second, the 'closer' the assumed to the 'true' correlation structure, the higher the value of pseudo-R^2_T, where pseudo-$R^2_T \approx R^2_{0T}$ if the assumed and the 'true' correlation structures coincide. The differences are more pronounced for a medium 'fit', whereas they are negligible for a very high or very low 'fit'. Third, there is not much difference between the various sample sizes, type of covariates used and the different number of observation points within each unit. Fourth, given a correlation structure, the larger the values of the correlations, the larger the differences between the values of pseudo-R^2_T for different estimated correlation structures.

5 Concluding Remarks

The simulation results suggest that the pseudo-R^2_T measure as proposed by Spiess and Keller (1999) represents the goodness-of-fit of the underlying model in the sense of taking on higher values if the error variance decreases, with everything else being constant.

An alternative to using pseudo-R^2_T would be to calculate pseudo-R^2, proposed by McKelvey and Zavoina (1975). If observations within subjects are assumed to be uncorrelated, then pseudo-R^2_T and pseudo-R^2 coincide. However, if observations within subjects are assumed to be correlated, then given appropriate

regression parameter estimates, pseudo-R^2_T should be used since it explicitely accounts for these correlations. In the latter case, the interpretation of pseudo-R^2 is questionable, since it does not represent the goodness-of-fit of the estimated model.

Beside the simplicity of its calculation, the advantage of using pseudo-R^2_T is that it allows to differentiate between various assumed correlation structures. These differences are most prominent in situations where observations within subjects are correlated and there is neither a perfect 'fit' nor no 'fit' at all of the systematic part of the model. Therefore, it may be concluded that, similar to R^2, R^2_T or pseudo-R^2, the measure pseudo-R^2_T may be used as an approximate descriptive measure of the goodness-of-fit of the estimated categorical panel probit model. In interpreting the value of pseudo-R^2_T, however, it must be kept in mind that unlike in linear uni- or multivariate models, the total SSP matrix cannot be interpreted as the sum of two orthogonal components, the residual and the fitted SSP matrix, in finite samples, even in linear panel models.

References

Hagle, T.M. & Mitchell II, G.E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science,* **36**, 762–784.

Hooper, J.W. (1959). Simultaneous equations and canonical correlation theory. *Econometrica,* **27**, 245–256.

McKelvey, R.D. & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology,* **4**, 103–120.

Spiess, M. (1998). A mixed approach for the estimation of probit models with correlated responses: Some finite sample results. *Journal of Statistical Computation and Simulation,* **61**, 39–59.

Spiess, M. & Keller, F. (1999). A mixed approach and a distribution-free multiple imputation technique for the estimation of a multivariate probit model with missing values. *British Journal of Mathematical and Statistical Psychology,* **52**, 1–17.

Veall, M.R. & Zimmermann, K.F. (1992). Pseudo-R^2's in the ordinal probit model. *Journal of Mathematical Sociology,* **16**, 333–342.

Veall, M.R. & Zimmermann, K.F. (1996). Pseudo-R^2 measures for some common limited dependent variable models. *Journal of Economic Surveys,* **10(3)**, 241–259.

Windmeijer, F.A.G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews,* **14(1)**, 101–116.

Figure 1: *'True' R^2_{0T} vs. pseudo-R^2_T using different estimation models (Independence, Equicorrelation, AR(1), Toeplitz) for various 'true' models with an AR(1) ('$\rho = .5$' or '$\rho = .8$') or a free, Toeplitz-like ('free') correlation structure*