cquad: An R and Stata Package for Conditional Maximum Likelihood Estimation of Dynamic Binary Panel Data Models

Bartolucci, Francesco and Pigini, Claudia

University of Perugia, University of Ancona

2 October 2015

`cquad`: An `R` and `Stata` Package for Conditional Maximum Likelihood Estimation of Dynamic Binary Panel Data Models

Francesco Bartolucci*
Università of Perugia

Claudia Pigini
Università Politecnica delle Marche

October 4, 2015

Abstract

We illustrate `R` package `cquad` for conditional maximum likelihood estimation of the quadratic exponential (QE) model proposed by Bartolucci and Nigro (2010) for the analysis of binary panel data. The package also allows us to estimate certain modified versions of the QE model, which are based on alternative parametrizations, and it includes a function for the pseudo conditional likelihood estimation of the dynamic logit model, as proposed by Bartolucci and Nigro (2012). We also illustrate a reduced version of this package that is available in `Stata`. The use of the main functions of this package is based on examples using labor market data.

KEYWORDS: dynamic logit model, pseudo maximum likelihood estimation, quadratic exponential model, state-dependence

1 Introduction

With the growing number of panel datasets available to practitioners and the recent development of related statistical and econometric models, ready-to-use software to estimate non-linear models for binary panel data is now

*Department of Economics, Università di Perugia, 06123 Perugia, Italia, email: francesco.bartolucci@unipg.it, URL: `https://sites.google.com/site/bartstatistics/`

essential in applied research. In particular, the panel structure allows for formulations that include both unobserved heterogeneity (i.e., time-constant individual intercepts) and the lagged response variable, which accounts for the so-called state dependence (i.e., how the experience of a certain event affects the probability of experiencing the same event in the future), as defined in Heckman (1981a).

A simple and, at the same time, interesting approach for the analysis of binary panel data is based on the dynamic logit (DL) model, which includes individual-specific intercepts and state dependence. The estimation of such a model may be based either on a random-effects or on a fixed-effects formulation. In the first case, individual intercepts are treated as random parameters while in the second each intercept is considered as a fixed parameter to be estimated. The fixed-effects approach attracts considerable more attention as it requires a reduced amount of assumptions with respect to the random-effects formulation, based on the independence between the individual unobserved effects and the observable covariates and the normality assumption.

For the static fixed-effects logit model (i.e., the DL without the lagged response among the covariates), it is possible to eliminate the individual intercepts by conditioning on simple sufficient statistics (Andersen, 1970; Chamberlain, 1980). In general, the estimator based on this method is known as Conditional Maximum Likelihood (CML) estimator. The general DL model, however, does not admit a simple sufficient statistic for the individual intercepts and, therefore, cannot be estimated by CML in a simple way as the static logit model.

The drawback described above is overcome by Bartolucci and Nigro (2010), who develop a model for the analysis of dynamic binary panel data models based on a Quadratic Exponential (QE) formulation (Cox, 1972), that has the advantage of admitting sufficient statistics for the unobserved heterogeneity parameters. The model parameters can therefore be easily estimated by the CML method. Recently, further extensions to Bartolucci and Nigro (2010) have also been proposed. In particular, Bartolucci and Nigro (2012) proposed a QE model that approximates more closely the DL model. Finally, Bartolucci et al. (2015) propose a test for state dependence that is more powerful than the one based on the standard QE model.

In this paper we illustrate `cquad`, available at `http://cran.r-project.org/web/packages/cquad/index.html`, which is a comprehensive `R` package for the CML estimation of fixed-effects binary panel data models. In particular, cquad contains functions for the estimation of the static logit model (Chamberlain, 1980), and of the dynamic QE models recently proposed by Bartolucci and Nigro (2010, 2012) and Bartolucci et al. (2015). A version of the R package cquad, including its main functionalities, is also available for Stata at https://ideas.repec.org/c/boc/bocode/s457891.html and is here illustrated.

As it implements fixed-effects estimators of non-linear panel data models for binary dependent variables, cquad complements the existing array of R packages for panel data econometrics. Above all, it closely relates to the plm package (see Croissant and Millo, 2008), which provides a wide set of functions for the estimation of linear panel data models, for both static and dynamic formulations. In addition, cquad shares with plm the peculiarities of the data frame structure, of the formula supplied to model.matrix, and of the object class panelmodel. cquad also relates to the package nlme (Pinheiro et al., 2015), which implements non-lineal mixed-effects models that can be estimated with longitudinal data.

The Stata module cquad represents an addition to the many existing commands and modules for panel data econometrics available in this software, such as xtreg and xtabond2 for linear model, and it complements the available routine for the CML and ML estimation of the static logit model, that is, the native xtlogit. In addition, it relates to the routines and modules for the estimation of random-effects binary panel data static models, such as the built-in xtprobit, the module gllamm (1999) for the estimation for generalized linear mixed models (see Rabe-Hesketh et al., 2005), and dynamic models, implemented in the modules redprob and redpace (see Stewart, 2006).

Finally, a package for the estimation of binary panel data models with similar functionalities is the DPB function package for Gretl (see Lucchetti and Pigini, 2015, for details), which implements the CML estimator for the QE model by Bartolucci and Nigro (2010). A related package, which however uses a different approach for parameter estimation, is the R package panelMPL described in Bartolucci et al. (2014).

The paper is organized as follows. In the next Section we briefly review the basic definition of the DL model and of the different version of QE models here considered. We also briefly review CML and pseudo-CML estimation of the models. Then, in Section 3 we describe the main functionalities of the package cquad for R and and the corresponding module for Stata. Finally, the illustration of the packages by examples is provided in Section 4.

For the purpose of describing `cquad` functionalities, we use data on unionized workers extracted from the U.S. National Longitudinal Survey of Youth. To illustrate the `R` package, we use the same data as in Wooldridge (2005) available at `http://qed.econ.queensu.ca/jae/datasets/wooldridge001/`, whereas for the `Stata` module we employ similar data already available in the `Stata` repository.

2 Preliminaries

We consider a binary panel dataset referred to a sample of n units observed at T consecutive time occasions. We adopt a common notation in which y_{it} is the response variable for unit i at occasion t, with $i = 1, \dots, n$ and $t = 1, \dots, T$, and $\boldsymbol{x}_{it}$ is the corresponding column of covariates. In the following we first describe the CML method applied to the logit model, then we illustrate the DL and QE models for the analysis of dynamic binary panel data models and inference based on the CML method.

2.1 Conditional maximum likelihood estimation

In order to provide an outline of the CML method by Andersen (1970), in the following we describe the derivation of the conditional likelihood for the static logit model (Chamberlain, 1980), which will serve as a basic framework for the QE models described later in this section.

Consider the static logit formulation based on the assumption

$$p(y_{it}|\alpha_i, \boldsymbol{X}_i) = \frac{\exp[y_{it}(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta})]}{1 + \exp(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta})},$$

where α_i is the individual specific intercept and vector $\boldsymbol{\beta}$ collects the regression parameters associated with the explanatory variables $\boldsymbol{x}_{it}$. For the joint probability of $\boldsymbol{y}_i$, this model implies that

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i) = \frac{\exp\left(\alpha_i y_{i+}\right) \exp\left[\left(\sum_t y_{it}\boldsymbol{x}_{it}\right)^\top \boldsymbol{\beta}\right]}{\prod_t \left[1 + \exp\left(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta}\right)\right]}, \tag{1}$$

where $y_{i+} = \sum_t y_{it}$ is called the *total score*.

It can be shown that y_{i+} is a sufficient statistic for the individual intercepts α_i (Andersen, 1970). Consequently, the joint probability of $\boldsymbol{y}_i$, conditional on y_{i+}, does not depend on α_i. In fact, we have

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i+}) = \frac{p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i)}{p(y_{i+}|\alpha_i, \boldsymbol{X}_i)},$$

where the denominator is the sum of the probabilities of observing each possible vector configuration of binary responses $\boldsymbol{z} = [z_1, \dots, z_T]^\top$ such that $z_+ = y_{i+}$, where $z_+ = \sum_t z_t$:

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i+}) = \frac{p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i)}{\sum\limits_{\boldsymbol{z}:z_+=y_{i+}} p(\boldsymbol{z}|\alpha_i, \boldsymbol{X}_i)},$$

with

$$p(\boldsymbol{z}|\alpha_i, \boldsymbol{X}_i) = \frac{\exp(\alpha_i z_+)\exp\left[\left(\sum_t z_t \boldsymbol{x}_{it}\right)^\top \boldsymbol{\beta}\right]}{\prod_t \left[1 + \exp\left(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta}\right)\right]}.$$

Therefore, the conditional distribution of the vector of responses $\boldsymbol{y}_i$ is

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i+}) = \frac{\exp(\alpha_i y_{i+})\exp\left[\left(\sum_t y_{it} \boldsymbol{x}_{it}\right)^\top \boldsymbol{\beta}\right]}{\prod_t \left[1 + \exp\left(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta}\right)\right]} \frac{\prod_t \left[1 + \exp\left(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta}\right)\right]}{\sum\limits_{\boldsymbol{z}:z_+=y_{i+}} \exp(\alpha_i z_+)\exp\left[\left(\sum_t z_+ \boldsymbol{x}_{it}\right)^\top \boldsymbol{\beta}\right]} \tag{2}$$

$$= \frac{\exp\left[\left(\sum_t y_{it} \boldsymbol{x}_{it}\right)^\top \boldsymbol{\beta}\right]}{\sum\limits_{\boldsymbol{z}:z_+=y_{i+}} \exp\left[\left(\sum_t z_t \boldsymbol{x}_{it}\right)^\top \boldsymbol{\beta}\right]} = p(\boldsymbol{y}_i|\alpha_i \boldsymbol{X}_i, y_{i+})$$

where the individual intercepts α_i have been canceled out.

The conditional log-likelihood based on the above distribution can be written as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{t} \mathbf{I}(0 < y_{i+} < T) \ln p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i+})$$

where the indicator function $\mathbf{I}(\cdot)$ is introduce to take into account that observations whose total score is 0 or T do not contribute to the likelihood. This conditional log-likelihood can be maximized with respect to $\boldsymbol{\beta}$ by the Newton-Raphson algorithm. Expressions for the score vector and information matrices can be derived using the standard theory on the regular exponential family (Barndorff-Nielsen, 1978).

2.2 Dynamic Logit model

The DL model (Hsiao, 2005) represents a interesting dynamic approach for binary panel data as, apart from the observable covariates, it includes both individual specific intercepts and the lagged response variable. Its formulation is a simple extension of Equation (1) with $y_{i,t-1}$ in the set of covariates.

For a sequence of binary responses y_{it}, $t = 1, \dots, T$, referred to the same unit i and the corresponding covariate vectors $\boldsymbol{x}_{it}$, the conditional distribution of a single response is

$$p(y_{it}|\alpha_i, \boldsymbol{X}_i, y_{i0}, \dots, y_{i,t-1}) = \frac{\exp[y_{it}(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + y_{i,t-1}\gamma)]}{1 + \exp(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + y_{i,t-1}\gamma)}$$

where γ is the regression coefficient for the lagged response variable which measures the true state dependence.

The inclusion of the individual intercept α_i for the unobserved heterogeneity in a dynamic model raises the so-called “initial conditions” problem (Heckman, 1981b), that concerns the correlation between time-invariant effects and the initial realization of the outcome y_{i0}. However, with a fixed-effects approach, individual unobserved effects are treated as fixed parameters and the initial observation can be considered as given. The distribution of the vector of responses $\boldsymbol{y}_i = (y_{i1}, \dots, y_{iT})^\top$ conditional on y_{i0} is

$$p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp\left(y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}_{it}^\top \boldsymbol{\beta} + y_{i*}\gamma\right)}{\prod_t \left[1 + \exp\left(\alpha_i + \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + y_{i,t-1}\gamma\right)\right]}, \tag{3}$$

where the sum $\sum_t$ and product $\prod_t$ range over $t = 1 \dots T$ and $y_{i*} = \sum_t y_{i,t-1}y_{it}$.

Differently from the static fixed-effects logit model in Equation (1), the DL model does not admit the total score as a sufficient statistic for the individual parameters α_i. Therefore, CML inference is not viable in a simple form, but can only be derived in the special case of $T = 3$ and in absence of explanatory variables (Chamberlain, 1985). Honoré and Kyriazidou (2000) extended this approach to include covariates in the regression model, so that parameters are estimated by CML on the basis of a weighted conditional log-likelihood. However, their approach presents some limitations, mainly discrete covariates cannot be included in the model specification and the rate of convergence is slower than $\sqrt{n}$, although the estimator is consistent.

2.3 Quadratic exponential models

The shortcomings of the fixed-effects DL model can be overcome by an approximating QE model defined in Bartolucci and Nigro (2010), based on the family of distributions for multivariate binary data formulated by Cox (1972). The QE_ext model directly formulates the conditional distribution of $\boldsymbol{y}_i$ as follows:

$$p(\boldsymbol{y}_i|\delta_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp\left[y_{i+}\delta_i + \sum_t y_{it}\boldsymbol{x}_{it}^\top\boldsymbol{\eta}_1 + y_{iT}\left(\phi + \boldsymbol{x}_{iT}^\top\boldsymbol{\eta}_2\right) + y_{i*}\psi\right]}{\sum_{\boldsymbol{z}} \exp[z_+\delta_i + \sum_t z_t\boldsymbol{x}_{it}^\top\boldsymbol{\eta}_1 + z_T\left(\phi + \boldsymbol{x}_{iT}^\top\boldsymbol{\eta}_2\right) + z_{i*}\psi]},$$

where δ_i is the individual specific intercept, $\sum_{\boldsymbol{z}}$ ranges over the possible binary response vectors $\boldsymbol{z}$, and $z_{i*} = y_{i0}z_1 + \sum_{t>1} z_{t-1}z_t$. The parameter ψ measures the true state dependence and vector $\boldsymbol{\eta}_1$ collects the regression parameters associated with the covariates. Here we consider ϕ and $\boldsymbol{\eta}_2$ as nuisance parameters. We refer the reader to Bartolucci and Nigro (2010) for the discussion on the interpretation of these parameters.

The QE model allows for state dependence and unobserved heterogeneity, other than the effect of observable covariates, some of which may be also discrete. Moreover, it shares several properties with the DL model:

1. for $t = 2, \ldots, T$, y_{it} is conditionally independent of $y_{i0}, \ldots, y_{i,t-2}$, given $\boldsymbol{X}_i, y_{i,t-1}$, and α_i or δ_i, under both models;

2. for $t = 1, \ldots, T$, the conditional log-odds ratio for $(y_{i,t-1}, y_{it})$ is constant:

$$\log\frac{p(y_{it} = 1|\delta_i, \boldsymbol{X}_i, y_{i,t-1} = 1)p(y_{it} = 0|\delta_i, \boldsymbol{X}_i, y_{i,t-1} = 0)}{p(y_{it} = 0|\delta_i, \boldsymbol{X}_i, y_{i,t-1} = 1)p(y_{it} = 1|\delta_i, \boldsymbol{X}_i, y_{i,t-1} = 0)} = \psi,$$

while in the DL model is constant and equal to γ.

Differently from the DL model, the QE model does admit a sufficient statistic for the individual intercepts δ_i. The parameters for the unobserved heterogeneity are removed by condition on the total score y_{i+}, by the same derivation in Section 2.1, so as to obtain the following expression:

$$p(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+}) = \frac{\exp[\sum_t y_{it}\boldsymbol{x}_{it}^\top\boldsymbol{\eta}_1 + y_{iT}(\phi + \boldsymbol{x}_{iT}^\top\boldsymbol{\eta}_2) + y_{i*}\psi]}{\sum_{\boldsymbol{z}:z_+=y_{i+}} \exp[\sum_t z_t\boldsymbol{x}_{it}^\top\boldsymbol{\eta}_1 + z_T(\phi + \boldsymbol{x}_{iT}^\top\boldsymbol{\eta}_2) + z_{i*}\psi]}. \quad (4)$$

The parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\eta}_1^\top, \phi, \boldsymbol{\eta}_2^\top, \psi\right)^\top$ can be estimated by CML by maximizing the conditional log-likelihood based on Equation (4), that is,

$$\ell(\boldsymbol{\theta}) = \sum_i \mathbf{I}(0 < y_{i+} < T) \log p(\boldsymbol{y}_i | \boldsymbol{X}_i, y_{i0}, y_{i+}).$$

As for the static logit model, the maximization may simply be performed by Newton-Raphson, and the resulting estimator is $\sqrt{n}$-consistent and has asymptotic normal distribution. For the derivation of the score and Information matrices and of the expression of the standard errors, we refer the reader to Bartolucci and Nigro (2010).

A simplified version of the QE_ext model can be derived by assuming that the regression parameters are equal for all time occasions. The joint probability of the individual outcomes of this model, which we will refer to as QE_basic henceforth, is expressed as

$$p^*(\boldsymbol{y}_i | \boldsymbol{X}_i, y_{i0}, y_{i+}) = \frac{\exp(\sum_t y_{it} \boldsymbol{x}_{it}^\top \boldsymbol{\phi} + y_{i*}\psi)}{\sum\limits_{\boldsymbol{z}: z_+ = y_{i+}} \exp(\sum_t z_t \boldsymbol{x}_{it}^\top \boldsymbol{\phi} + z_{i*}\psi)}. \tag{5}$$

In the same way as for the QE_ext model, a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta} = \left(\boldsymbol{\phi}^\top, \psi\right)^\top$ can be obtained by maximizing the conditional log-likelihood based on (5) by Newton-Raphson.

Finally, Bartolucci et al. (2015) introduce a test for state dependence based on a modified version of the QE_basic model, named QE_equ hereafter. The joint probability of $\boldsymbol{y}_i$ is defined as

$$\tilde{p}(\boldsymbol{y}_i | \delta_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp(y_{i+}\delta_i + \sum_t y_{it} \boldsymbol{x}_{it}^\top \boldsymbol{\phi} + \tilde{y}_{i*}\psi)}{\sum_{\boldsymbol{z}} \exp(z_+\delta_i + \sum_t z_t \boldsymbol{x}_{it}^\top \boldsymbol{\phi} + \tilde{z}_{i*}\psi)},$$

where $\tilde{y}_{i*} = \sum_t \mathbf{I}\{y_{it} = y_{i,t-1}\}$ and $\tilde{z}_{i*} = \mathbf{I}\{z_1 = y_{i0}\} + \sum_{t>1} \mathbf{I}\{z_t = z_{t-1}\}$. The difference with the QE models described earlier is in how the association between the response variables is formulated: this modified version is based on the statistic $\tilde{y}_{i*}$ that, differently from y_{i*}, is equal to the number of consecutive pairs of outcomes which are equal each other, regardless if they are 0 or 1. This allows us to use a larger set of information with respect to the QE_ext and QE_basic in testing for state dependence.

Conditioning on the total score y_{i+}, the expression for the joint probability becomes

$$\tilde{p}(\boldsymbol{y}_i | \boldsymbol{X}_i, y_{i0}, y_{i+}) = \frac{\exp(\sum_t y_{it} \boldsymbol{x}_{it}^\top \boldsymbol{\phi} + \tilde{y}_{i*}\psi)}{\sum\limits_{\boldsymbol{z}: z_+ = y_{i+}} \exp(\sum_t z_t \boldsymbol{x}_{it}^\top \boldsymbol{\phi} + \tilde{z}_{i*}\psi)}. \tag{6}$$

In the same way as for the QE_ext and QE_basic model, $\boldsymbol{\theta} = \left(\boldsymbol{\phi}^\top, \psi\right)^\top$ can be consistently estimated by CML and, in particular, by maximizing the conditional log-likelihood based on (6).

Once the parameters in Equation (6) are estimated, a t-statistic for $H_0 : \psi = 0$ is

$$W = \frac{\tilde{\psi}}{\mathrm{se}(\tilde{\psi})}, \tag{7}$$

where $\mathrm{se}(\cdot)$ is the standard error derived using the sandwich estimator; see Bartolucci et al. (2015) for the complete derivation of score, information matrix, and variance-covariance matrix.

Under the DL model, and provided that the null hypothesis $H_0 : \gamma = 0$ holds, the test statistic W has asymptotic standard normal distribution as $n \to \infty$. If $\gamma \neq 0$, W diverges to $+\infty$ or $-\infty$ according to whether γ is positive or negative.

2.4 Pseudo-conditional maximum likelihood estimation

In order to estimate the structural parameters of the DL model, Bartolucci and Nigro (2012) propose a pseudo-CML estimator based on approximating the DL model by a QE model of the type described in Section 2.3. The proposed approximating model also has the advantage of admitting a simple sufficient statistic for the individual intercepts and its parameters share the same interpretation with those of the true DL model.

The approximating model is derived from a linearization of the log-probability of the DL model defined in Equation (3), that is,

$$\log[p(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0})] = y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}_{it}^\top\boldsymbol{\beta} + y_{i*}\gamma - \sum_t \log[1+\exp(\alpha_i + \boldsymbol{x}_{it}^\top\boldsymbol{\beta} + y_{i,t-1}\gamma)].$$

The non-linear component is approximated by a first-order Taylor series expansion around $\alpha_i = \bar{\alpha}$, $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$, and $\gamma = 0$:

$$\begin{aligned}\sum_t \log[1 + \exp(\alpha_i + \boldsymbol{x}_{it}^\top\boldsymbol{\beta} + y_{i,t-1}\gamma)] \quad &\approx \quad \sum_t \{\log\left[1 + \exp\left(\bar{\alpha}_i + \boldsymbol{x}_{it}^\top\bar{\boldsymbol{\beta}}\right)\right] + \\ +\bar{q}_{i1}\left[\alpha_i - \bar{\alpha}_i + \boldsymbol{x}_{it}^\top(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right]\} \quad &+ \quad \bar{q}_{i1}y_{i0}\gamma + \sum_{t>1}\bar{q}_{it}y_{i,t-1}\gamma,\end{aligned}$$

where $\bar{q}_{it} = \exp(\bar{\alpha}_i + \boldsymbol{x}_{it}^\top \bar{\boldsymbol{\beta}})/[1+\exp(\bar{\alpha}_i + \boldsymbol{x}_{it}^\top \bar{\boldsymbol{\beta}})]$. Under this approximating model, referred to QE_pseudo hereafter, the joint probability of $\boldsymbol{y}_i$ is

$$p^\dagger(\boldsymbol{y}_i|\alpha_i, \boldsymbol{X}_i, y_{i0}) = \frac{\exp(y_{i+}\alpha_i + \sum_t y_{it}\boldsymbol{x}_{it}^\top \boldsymbol{\beta} - \sum_t \bar{q}_{it} y_{i,t-1}\gamma + y_{i*}\gamma)}{\sum_{\boldsymbol{z}} \exp(z_+\alpha_i + \sum_t z_t \boldsymbol{x}_{it}^\top \boldsymbol{\beta} - \sum_t \bar{q}_{it} z_{i,t-1}\gamma + z_{i*}\gamma)}.$$

Given α_i and $\boldsymbol{X}_i$, the above model corresponds to a quadratic exponential model (Cox, 1972) with second-order interactions equal to γ, when referred to consecutive response variables, and to 0 otherwise.

Under the approximating model, each y_{i+} is a sufficient statistic for the incidental parameter α_i. By conditioning on the total scores, the joint probability of $\boldsymbol{y}_i$ becomes:

$$p^\dagger(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+}) = \frac{\exp(\sum_t y_{it}\boldsymbol{x}_{it}^\top \boldsymbol{\beta} - \sum_t \bar{q}_{it} y_{i,t-1}\gamma + y_{i*}\gamma)}{\sum\limits_{\boldsymbol{z}: z_+ = y_{i+}} \exp(\sum_t z_t \boldsymbol{x}_{it}^\top \boldsymbol{\beta} - \sum_t \bar{q}_{it} z_{i,t-1}\gamma + z_{i*}\gamma)}, \quad (8)$$

where the individual intercepts α_i have been canceled out.

A pseudo-CML estimator based on the approximating model described in Equation (8) is introduced by Bartolucci and Nigro (2012). The estimator is based relies on the following two-step procedure:

1. A preliminary estimate of the regression parameter $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$, is computed by maximizing the conditional log-likelihood of the static logit model described in Section 2.1. In addition, the probabilities $\bar{q}_{it}$, $i = 1, \ldots, n$, $t = 2, \ldots, T$ are computed in $\bar{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ and $\bar{\alpha}_i$ equal to its maximum likelihood estimate under the static logit model.

2. The parameter vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \gamma\right)^\top$ is estimated by maximizing the conditional log-likelihood

$$\ell^\dagger(\boldsymbol{\theta}|\bar{\boldsymbol{\beta}}) = \sum_i \mathbf{I}\{0 < y_{i+} < T\}\ell_i^\dagger(\boldsymbol{\theta}|\bar{\boldsymbol{\beta}}),$$

where

$$\ell_i^\dagger(\boldsymbol{\theta}|\bar{\boldsymbol{\beta}}) = \log[p^\dagger(\boldsymbol{y}_i|\boldsymbol{X}_i, y_{i0}, y_{i+})].$$

Maximization of $\ell^\dagger(\boldsymbol{\theta}|\bar{\boldsymbol{\beta}})$ is possible by a simple Newton-Raphson algorithm, resulting in the pseudo-CML estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\gamma})^\top$ of the structural parameters of the DL model. For asymptotic results and computation of standard errors we refer the reader to Bartolucci and Nigro (2012).

3 Package description

Here we describe the main functionalities of the R package cquad and then the corresponding commands of the cquad module implemented in Stata.

3.1 The R package

3.1.1 The cquad interface

The package cquad includes several functions, the majority of which are called by the main interface cquad. The first argument of the cquad function is a formula that shares the same syntax with that of the plm package. For instance, using the sample data on unionized workers, Union.RData, a simple function call is

```
cquad(union ~ married, Union)
```

where the dependent variable must be a numeric binary vector. In general, as in plm and differently from lm, the formula can also recognize the operators lag, log and diff that can be supplied directly without additional transformations of the covariates.

The second argument supplied to cquad is the data frame. As in plm, the data must have a panel structure, that is the data frame has to contain an individual identifier and a time variable as the first two columns. For instance, the data frame Union has the following structure:

```
head(Union[c(1,2)])

  nr year
1 13 1980
2 13 1981
3 13 1982
4 13 1983
5 13 1984
6 13 1985
```

where nr is the individual identifier and year provides the time variable. As Union already has a panel structure, cquad can be called directly. Differently, if the dataset does not contain the individual and time indicators, cquad can set the panel structure and create automatically the first two variables, provided the index is supplied, that is the number of cross-section observations in the data. As an example, the dataset `Wages`, supplied by `plm` and containing 595 individuals observed over 7 periods, does not have a panel structure, which can however be imposed by `cquad` as follows

```
cquad(union2 ~ married, Wages, index = 595)
```

The package `cquad` uses the same function as `plm` to impose the panel structure on a data frame, called `plm.data`. Indeed, this function can also be used to set the structure to the data frame that can then be supplied to `cquad` without the index argument. For instance:

```
 Wages = plm.data(Wages, 595)
```

produces

```
 head(Wages)
```

```
  id time exp wks bluecol ind south smsa married  sex union ed black   lwage
1  1    1   3  32      no   0   yes   no     yes male    no  9    no 5.56068
2  1    2   4  43      no   0   yes   no     yes male    no  9    no 5.72031
3  1    3   5  40      no   0   yes   no     yes male    no  9    no 5.99645
4  1    4   6  39      no   0   yes   no     yes male    no  9    no 5.99645
5  1    5   7  42      no   1   yes   no     yes male    no  9    no 6.06146
6  1    6   8  35      no   1   yes   no     yes male    no  9    no 6.17379
```

where the factors `id` and `time` have been created and added to the data frame.

In the examples above, both data frames contain balanced datasets. Nevertheless, `cquad` also handles unbalanced panels.

Each of the models described in Section 2 are estimated by `cquad` by supplying a dedicated string to the function argument `model`. Specifically:

- The fixed-effects static logit model by Chamberlain (1980) (`model = "basic"`, default);
- The simplified QE model, QE_basic (`model = "basic", dyn = TRUE`);
- The QE_ext model proposed by Bartolucci and Nigro (2010) (`model = "extended"`);
- The modified version of the QE model, QE_equ proposed in Bartolucci et al. (2015) (`model = "equal"`);

- The pseudo-CML estimation of the DL model based on the approach of Bartolucci and Nigro (2012) (`model = "pseudo"`).

As an optional argument, the `cquad` function can also be supplied with a n-vector of individual weights, the default value is `rep(1, n)`.

The results of the calls to `cquad` are stored in an object of class `panelmodel`. The returned object shares only some elements with a `panelmodel` object and contains additional ones due to the peculiarities of CML inference.

The elements in common with the object `panelmodel` as described in `plm` are `coefficients`, `vcov`, and `call`. The vector `coefficients` contains the estimates of: the k-vector $\boldsymbol{\beta}$, for the static logit; the $k+1$-vector $\boldsymbol{\theta} = \left(\boldsymbol{\phi}^\top, \psi\right)^\top$ for the dynamic models QE_basic and QE_equ in Equation (5) and Equation (6), respectively; the $2k+2$-vector for the QE_ext model, where the vector consists of $\left(\boldsymbol{\eta}_1^\top, \phi, \boldsymbol{\eta}_2^\top, \psi\right)^\top$ in Equation (4); the $k+1$-vector $\left(\boldsymbol{\beta}^\top, \gamma\right)^\top$ in Equation (8) for the pseudo-CML estimator of the DL model. The matrix `vcov` contains the corresponding asymptotic variance-covariance matrix for the parameter estimates. Finally, `call` contains the function call to the sub-routines required to fit each model, namely `cquad_basic`, `cquad_ext`, `cquad_equ`, and `cquad_pseudo`.

The output of `cquad` does not provide fitted values nor residuals: as discussed in Section 2, the CML estimation approach is based on eliminating the individual intercepts in each model, which does not allow for the computation of predicted probabilities. Similarly, residuals are not a viable tool for standard inference. Instead, we supply the object with estimated quantities useful for inference and diagnostics in the CML estimation approach.

The asymptotic standard errors associated with the estimated coefficients are collected in the vector `se` and the robust standard errors (White, 1980) in vector `ser`. For the pseudo-CML estimator, the standard errors contained in the vector `ser` are corrected for the presence of generated regressors (see Bartolucci and Nigro, 2012, for the detailed derivation of the two-step variance-covariance matrix). The function output also provides the matrix `scv` containing the individual scores and the matrix `J` containing the Hessian of the log-likelihood function. In addition, `cquad` returns the conditional log-likelihood at convergence `lk` for each of the fitted models. Finally, it contains the n-vector `Tv` of the number of time occasions at which each unit is observed.

3.1.2 Simulate data from the DL model

The package `cquad` also provides the user with function `sim_panel_logit` to generate a binary vector from a DL data generating process. This function requires in input the list of identifiers of the dataset collected in vector `id` that has length equal to the overall number of observations $n \times T = r$. As other inputs, the function requires the n-dimensional vector of the individual specific intercepts that must be somehow generated, for instance drawing them from a standard normal distribution, and the matrix of covariates (if they exist) that has dimension $r \times k$, where k is the number of covariates. Each row of this matrix contains a vector of covariates $\boldsymbol{x}_{it}$ arranged according to vector `id`. Finally, in input the function requires the vector of structural parameters, denoted by `eta`, that is, $\boldsymbol{\beta}$ for the static logit model and $(\boldsymbol{\beta}^\top, \gamma)^\top$ for the DL model; the model of interest is specified by the optional argument `dyn`.

As output values, function `sim_panel_logit` returns a list containing two vectors, `pv` and `yv`. The first contains the success probability computed according to the DL model corresponding to each row of matrix `X` and accounting for the corresponding individual intercept in `al`. Vector `yv` contains the binary variable which is randomly drawn from this distribution.

3.2 The `Stata` module

The `cquad` module in `Stata` consists of four `mata` routines for the estimation by CML of the QE models described in Section 2.3. It contains four commands with the syntax

`cquad`*cmd* *depvar* `id` [*indepvars*]

where *cmd* has to be substituted with the string corresponding to the type of model to be estimated. In particular:

- `cquadext` fits the QE_ext model of Bartolucci and Nigro (2010) defined by the conditional probability in Equation (4);
- `cquadbasic` estimates the parameters of the simplified QE model, QE_basic, defined in Equation (5). Differently from the `R` package, `cquadbasic` fits only the dynamic QE model, as the static logit model can estimated by `xtlogit`;

- `cquadequ` fits the modified QE model defined by the conditional probability in Equation (6) proposed by Bartolucci et al. (2015);

- `cquadpseudo` fits the pseudo-CML estimator proposed by Bartolucci and Nigro (2012) for the parameters in Equation (8).

In addition, *depvar* is the series containing the binary dependent variable, and `id` is the variable containing the list of reference unit uniquely identifying individuals in the panel dataset. Optionally a list of covariates [*indepvars*] can be supplied.

The four commands return an `eclass` object with the estimation results. Scalar `e(lk)` contains the final conditional log-likelihood, macro `e(cmd)` holds the function call. Matrix `e(be)` contains the estimated coefficients and it is of dimension $(2k+2)\times 1$ after `cquadext`, or of dimension $(k+1)\times 1$ after `cquadbasic`, `cquadequ` and `cquadpseudo`. Matrices `e(se)` and `e(ser)` contain the corresponding estimated asymptotic and robust standard errors, respectively. Finally, matrices `e(tstat)` and `e(pv)` collect the t-tests and the corresponding p-values.

4 Examples

In the following we illustrate `cquad` in `R` and `Stata` by means of three applications. We show how to fit CML estimators for the QE models and the pseudo-CML estimator in `R` and `Stata` using longitudinal data on unionized workers extracted from the U.S. National Longitudinal Survey of Youth, which has been employed in several applied works to illustrate dynamic binary panel data models (Wooldridge, 2005; Stewart, 2006; Lucchetti and Pigini, 2015). In addition, we propose a simulation example using `sim_panel_logit` provided by the `R` package.

4.1 Use of the `Union` dataset in `R`

To illustrate the `R`, we use the dataset employed in Wooldridge (2005) and available in the Journal of Applied Econometrics data archive. The dataset consists of 545 male workers interviewed for eight years, from 1980 to 1987. Similarly to the empirical application Wooldridge (2005), the variables relevant to our example are a binary variable equal to 1 if the worker's wage is set by an union, which will be used as the dependent variable, and a binary variable describing his marital status, used as covariate. The original datatset also contains information on the men race and years of schooling, which however cannot be employed in our example since they are time-invariant:

```
  nr year black married educ union
1 13 1980     0       0   14     0
2 13 1981     0       0   14     1
3 13 1982     0       0   14     0
4 13 1983     0       0   14     0
5 13 1984     0       0   14     0
6 13 1985     0       0   14     0
```

Notice that the panel structure required by `cqaud` is already in place.

Then, in order to fit the static logit model to this data by the CML method, we call `cquad` with the following syntax

```
out1 = cquad(union ~ married + year, Union)
```

with estimates a logit model with `union` as dependent variable and `married` and time dummies as covariates, obtaining the following output

```
Balanced panel data
 |--------------|--------------|--------------|
 |   iteration  |      lk      |    lk-lko    |
 |--------------|--------------|--------------|
 |            1 |     -740.781 |          Inf |
 |            2 |      -732.45 |       8.3312 |
 |            3 |     -732.445 |   0.00539603 |
 |            4 |     -732.445 |  9.75388e-09 |
 |--------------|--------------|--------------|
```

Then, using command `summary(out1)`, we obtain:

```
Call:
cquad_basic(id = id, yv = yv, X = X, w = w, dyn = dyn)

Log-likelihood:
-732.4449
                 est.      s.e.       t-stat     p-value
married   0.298326773 0.1708112  1.746529038 0.080719066
year1981 -0.061754846 0.2061185 -0.299608423 0.764475859
year1982  0.000927442 0.2069901  0.004480611 0.996425002
```

```
year1983 -0.155186804 0.2117482 -0.732883615 0.463629417
year1984 -0.107846793 0.2137133 -0.504633157 0.613816517
year1985 -0.442338283 0.2189339 -2.020419690 0.043339873
year1986 -0.608785100 0.2222082 -2.739705640 0.006149423
year1987 -0.015457650 0.2180398 -0.070893720 0.943482341
```

The output of `summary` displays the function call, the value of the log-likelihood at convergence, and the estimated coefficients with the corresponding asymptotic standard errors and t-test results. Notice that including the variable `year` in the sets of covariates in the formula, resulted in `cquad` automatically including time dummies in the model specification, except for `year1980` for collinearity, even though the variable `year` is numeric in the original data frame:

```
str(Union$year)
```

```
int [1:4360] 1980 1981 1982 1983 1984 1985 1986 1987 1980 1981 ...
```

This happens because `cquad` recognizes the second variable in the data frame as the time variable, and with the call to `plm.data` and `model.matrix` the numeric time variable is transformed into a factor.

To estimate the dynamic specification of the QE_basic model, `cquad` needs to be called with the `dyn = TRUE` switch. In addition, as we are working with a balanced panel, an additional time dummy has to be excluded because the lag of the dependent variable is included in the conditioning set and the initial time occasion is lost. In this case, we perform this operation outside the `cquad` interface

```
year2 = Union$year
year2[year2==1980 | year2==1981] = 0
year2 = as.factor(year2)
out2 = cquad(union ~ married + year2, Union, dyn=TRUE)
summary(out2)
```

In the code above, we store the numeric time variable from the original data frame in `year2`, then we set the variable to 0 for two of its values, as we loose one time occasion because of the dynamic specification and one time effect because of collinearity of the remaining dummies. In order to estimate the model with time dummies, we need to convert `year2` into a factor as `cquad` will not recognize `year2` as the time variable since it is not in the data frame. If instead we leave `year` in the formula, a warning message is printed after convergence and the results are obtained using the generalized inverse of the Hessian matrix.

The estimation output produced by the above command lines is (iteration logs are omitted from the output below)

```
Call:
cquad_basic(id = id, yv = yv, X = X, w = w, dyn = dyn)

Log-likelihood:
-505.514
                 est.      s.e.     t-stat      p-value
married    0.13404719 0.1868762  0.7173047 0.4731861145
year21982  0.09160286 0.2441350  0.3752140 0.7075013011
year21983 -0.09896744 0.2258889 -0.4381245 0.6612960556
year21984  0.09917729 0.2254660  0.4398770 0.6600262259
year21985 -0.27210110 0.2309277 -1.1782956 0.2386787776
year21986 -0.52465221 0.2328383 -2.2532900 0.0242408710
year21987  0.81055556 0.2265106  3.5784449 0.0003456447
y_lag      1.47082575 0.1528797  9.6208037 0.0000000000
```

Although `cquad` with `model = "basic"` (default) and `dyn = TRUE` fits the simplified version of the QE model, QE_basic, which approximates the true DL model, the obtained results are in line with the findings on the probability of participating in a union for dynamic models: there is a positive and significant correlation with the lagged dependent variable ($\psi = 1.471$), and the effect of `married` is not statistically significant.

To fit the QE_ext model, we need to further exclude the last time value 1987: since there is an intercept term ϕ in Equation (4), the effect related to the last time dummy is not identified with balanced panels:

```
year3 = Union$year
year3[year3==1980 | year3==1981 | year3==1987] = 0
year3 = as.factor(year3)
out3 = cquad(union ~ married + year3, Union, model = "extended")
```

By typing `summary(out3)` we obtain

```
Call:
cquad_ext(id = id, yv = yv, X = X, w = w)

Log-likelihood:
```

```
-504.2864
                    est.      s.e.      t-stat    p-value
married       0.01958449 0.2008834  0.09749182 0.92233583
year31982     0.09808421 0.2442447  0.40158167 0.68799192
year31983    -0.08051308 0.2262232 -0.35590102 0.72191469
year31984     0.12301583 0.2259423  0.54445680 0.58612717
year31985    -0.24494702 0.2314885 -1.05813907 0.28999205
year31986    -0.48914076 0.2339525 -2.09076982 0.03654870
int           0.51995850 0.2952783  1.76091005 0.07825363
diff.married  0.51942916 0.3328688  1.56046215 0.11865071
y_lag         1.47056206 0.1530829  9.60631199 0.00000000
```

where the additional `int` and `diff.` variables represent ϕ and $\boldsymbol{\eta}_2$ in Equation (4), respectively.

Similarly, to fit the QE_equ model defined in Equation (6) and display the results, the command line is as follows:

```
out4 = cquad(union ~ married + year2, Union, model = "equal")
summary(out4)
```

which returns

```
Call:
cquad_equ(id = id, yv = yv, X = X, w = w)

Log-likelihood:
-505.514
                 est.       s.e.     t-stat    p-value
married    0.13404719 0.18687622  0.7173047 0.47318611
year21982  0.09160286 0.24413496  0.3752140 0.70750130
year21983 -0.09896744 0.22588886 -0.4381245 0.66129606
year21984  0.09917729 0.22546598  0.4398770 0.66002623
year21985 -0.27210110 0.23092771 -1.1782956 0.23867878
year21986 -0.52465221 0.23283830 -2.2532900 0.02424087
year21987  0.07514269 0.21352948  0.3519078 0.72490741
y_lag      0.73541287 0.07643986  9.6208037 0.00000000
```

Notice that there is a marked different in the estimate corresponding to the lagged dependent variable. In model QE_equ, the association between y_{it} and $y_{i,t-1}$ is different from that of the standard formulation of the QE model in order to exploit more information in testing for state dependence (see Section 2.3). Indeed, the t-stat associated with `y_lag` reports the test for state dependence described in Equation (7).

In order to fit the pseudo-CML model, `cquad` needs to be called with `model = "pseudo"`:

```
out5 = cquad(union ~ married + year2, Union, model = "pseudo")
```

that produces the output

```
First step estimation
Balanced panel data
 |--------------|--------------|--------------|
 |   iteration  |      lk      |    lk-lko    |
 |--------------|--------------|--------------|
 |            1 |     -740.781 |          Inf |
 |            2 |     -732.495 |      8.28629 |
 |            3 |      -732.49 |   0.00541045 |
 |            4 |      -732.49 |   9.8679e-09 |
 |--------------|--------------|--------------|
Second step estimation
 |--------------|--------------|--------------|
 |   iteration  |      lk      |    lk-lko    |
 |--------------|--------------|--------------|
 |            1 |     -552.702 |          Inf |
 |            2 |     -528.266 |      24.4361 |
 |            3 |     -513.702 |      14.5641 |
 |            4 |     -509.195 |      4.50721 |
 |            5 |     -509.192 |   0.00285414 |
 |            6 |     -509.192 |  1.11389e-08 |
 |--------------|--------------|--------------|
```

where the first panel reports the iterations of the first step CML estimation of the regression coefficients in the static logit model, while the second refers to the second step maximization to obtain the pseudo-CML estimates of the parameters in Equation (8).

After calling `summary(out5)` the following results are displayed:

```
Call:
cquad_pseudo(id = id, yv = yv, X = X)
```

```
Log-likelihood:
-509.1917
                  est.      s.e.     t-stat      p-value
married    0.19259731 0.1858896  1.0360844 3.001628e-01
year21982  0.05031661 0.2664274  0.1888567 8.502051e-01
year21983 -0.12381494 0.2092980 -0.5915724 5.541369e-01
year21984 -0.02956563 0.2224643 -0.1329006 8.942720e-01
year21985 -0.43257573 0.2243302 -1.9282989 5.381796e-02
year21986 -0.54727988 0.2212247 -2.4738647 1.336603e-02
year21987  0.17223711 0.2425840  0.7100103 4.776978e-01
y_lag      1.47526322 0.1807924  8.1599843 4.440892e-16
```

Notice that the estimation results are coherent with those obtained by fitting the QE_ext or the QE_basic models, however they exhibit some differences since the pseudo-CML estimator is based on the conditional probability in Equation (8) that contains the parameters of the true DL model. Nevertheless, these results confirm the presence of a high degree of state dependence in union participation.

4.2 Use of `sim_panel_logit` to generate dynamic binary panel data

In the following, we illustrate how to perform a simple simulation study on data generated from a DL model by means of function `sim_panel_logit` in the package `cquad`. In this example, we fit the modified QE_equ model by CML and study the properties of the test for state dependence proposed by Bartolucci et al. (2015). The script to replicate the exercise is reported below

```
require(cquad)
# set simulation parameters
n = 500
TT = 6
nit = 100
be = 1
rho = 0.5
var = pi*pi^2/3
stdep = c(0,1)
TEST = rep(0,nit)
```

```
for(ga in stdep)
    for(it in 1:nit)

        # generate data
        label = 1:n
        id = rep(label,each=TT)
        X = matrix(rep(0),n*TT,1); alpha = rep(0,n); eta = rep(0,n*TT)
        e = rnorm(n*TT)*sqrt(var*(1-rho^2))
        j = 0
        for(i in 1:n)
          j = j+1
          X[j] = rnorm(1)*sqrt(var)
          for(t in 2:TT)
            j = j+1
            X[j] = rho*X[j-1] + e[j]

          alpha[i] = (X[j-2] + X[j-1] + X[j])/3

        cat("sample n. ",it,"
n")
        data = sim_panel_logit(id,alpha,X,c(be,ga),dyn=TRUE)
        yv   = data$yv

        # estimate QE equal
        mod = cquad(yv ~ X, data.frame(yv, X), index = 500, model = "equal")

        # store results
        beta = mod$coefficients
        TEST[it] = beta[length(beta)]/mod$se[length(beta)]


    # display results
    cat(c("gamma =", ga,"
n"))
    RES = c(mean(TEST), mean(abs(TEST)>1.96))
    names(RES) = c("t-stat", "rej. rate")
    print(RES)
```

In the first part of the script, we set the simulation parameters for the sample size, number of time occasions and number of Monte Carlo replications. We also set the parameter values for the DL model in Equation (3) with one regression parameter $\beta = 1$ and one covariate, generated as an AR(1) process with autocorrelation coefficient $\rho = 0.5$. In this exercise, we analyze two scenarios, with the state dependence parameter γ equal to 0 and 1.

In the first part of the script inside the `for` loops, we generate the identifier `id`, the covariate `X` and the n-vector of individual intercepts `alpha`, which is computed is a similar manner as in Honoré and Kyriazidou (2000). Lastly, we generate the binary response variable using function `sim_panel_logit` described in Section 3.1. As the function returns both the binary variable and the response probabilities, the dependent variable needs to be retrieved by `yv = data$yv`.

Once the data have been generated, we proceed to the estimation of the QE_equ model using `cquad` with `model = "equal"` to fit by CML the modified QE model in Equation (6) and we store the results for the t-test in Equation (7). Finally, we print the results where we display the average value of the test in the 100 sample and the average rejection rate of a bilateral test at the 0.05 significance level. The last part or the script produces the following output:

```
 ...

gamma = 0
    t-stat  rej. rate
-0.1753164  0.0400000

...

gamma = 1
   t-stat rej. rate
 4.939813  0.990000
```

where the iteration logs from `cquad` have been omitted. Under the null hypothesis of $\gamma = 0$, the rejection rate is very close to the nominal size of 0.05, while under the alternative of $\gamma = 1$ the test exhibits good power properties. These results are close to those found by Bartolucci et al. (2015) in their simulation study, to which we refer the reader for an extension of this simple design to several other scenarios.

4.3 Analysis of `union` data in `Stata`

In the following, we illustrate the `Stata` module `cquad` that contains the four commands to fit the QE models described in Section 2.3 by an example based on data on unionized workers, often employed to illustrate dynamic binary panel data models (Stewart, 2006; Lucchetti and Pigini, 2015). The dataset to replicate this example is available on the `Stata` online data repository as `union.dta`.

The three commands reported below load the dataset, then describe the panel structure, already in place, and list the variables present in the dataset

```
webuse union
xtdes
descr
```

The output generated by these command lines is:

```
. webuse union
(NLS Women 14-24 in 1968)

. xtdes

  idcode:  1, 2, ..., 5159                                n =       4434
    year:  70, 71, ..., 88                                T =         12
           Delta(year) = 1 unit
           Span(year)  = 19 periods
           (idcode*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%      50%      75%      95%      max
                         1       1       3        6        8       11       12

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------------------
      190      4.29    4.29 |  1111...11.1.11.1.11
      129      2.91    7.19 |  .......11.1.11.1.11
       93      2.10    9.29 |  1..................
       78      1.76   11.05 |  .......1...........
       68      1.53   12.58 |  ..11...11.1.11.1.11
       64      1.44   14.03 |  ...1...11.1.11.1.11
       60      1.35   15.38 |  .111...11.1.11.1.11
       52      1.17   16.55 |  11.................
```

```
      52      1.17   17.73 |  1111...............
    3648     82.27  100.00 | (other patterns)
 ---------------------------+---------------------
    4434    100.00         |  XXXX...XX.X.XX.X.XX
```

```
. descr

Contains data from http://www.stata-press.com/data/r13/union.dta
  obs:        26,200                          NLS Women 14-24 in 1968
 vars:             8                          4 May 2013 13:54
 size:       235,800
-------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------
idcode          int     %8.0g                 NLS ID
year            byte    %8.0g                 interview year
age             byte    %8.0g                 age in current year
grade           byte    %8.0g                 current grade completed
not_smsa        byte    %8.0g                 1 if not SMSA
south           byte    %8.0g                 1 if south
union           byte    %8.0g                 1 if union
black           byte    %8.0g                 race black
-------------------------------------------------------------------------
Sorted by:  idcode  year
```

The dataset consists of 4434 women that were between 14 and 24 years old in 1968, interviewed between 1970 and 1988. The panel is unbalanced and the maximum number of occasions of observation of the same subject is 12. The last part of the output reports the variable descriptions, where `union` will be the response variable in our exercise, `age`, `grade`, `not_smsa` and `south` the set of covariates, while `black` will be excluded from the analysis because of its time-invariant nature.

We first illustrate command `cquadbasic` to fit the QE_basic model in Equation (5) by CML, where we include time dummies in the model specification by using the `xi` and `i.year` declarations. The command line

```
xi: cquadbasic union idcode age grade south not_smsa i.year
```

produces the following output

```
. xi: cquadbasic union idcode age grade south not_smsa i.year
i.year           _Iyear_70-88       (naturally coded; _Iyear_70 omitted)

Fit (simplified) quadratic exponential model by Conditional Maximum Likelihood
see Bartolucci & Nigro (2010), Econometrica

      |      lk       lk-lk0
------+--------------------------------------------------
    1 | -3439.9096  1.000e+10
    2 | -3071.6412  368.26839
    3 | -3069.0539  2.5872579
    4 | -3069.0534  .00050444
    5 | -3069.0534  5.775e-11


             |      est.      s.e       t-stat.    p-value
-------------+--------------------------------------------------
         age |  .17670917   .1192216  1.4821908  .06914476
       grade | -.03658997  .04586492 -.79777692  .21249998
       south |  -.5191613  .13732314 -3.7805814  .00007823
    not_smsa |  .12631127  .13146408   .9608044  .16832526
   _Iyear_71 |  1.5208636   1.035464  1.4687749  .07094693
   _Iyear_72 |  1.1096837  .91812295  1.2086439  .11339984
   _Iyear_73 |  .90256541  .79733234  1.1319814  .12882112
   _Iyear_77 |   .1829554   .3308496  .55298662  .29013629
   _Iyear_78 |  .17904624  .21288676   .8410398  .20016282
   _Iyear_80 |  .46950024   .0846961  5.5433514  1.484e-08
   _Iyear_82 | -.40988205  .28664089 -1.4299497  .07636573
   _Iyear_83 | -.95438994  .40301052 -2.3681514  .00893861
   _Iyear_85 | -.73765258  .63762525 -1.1568748  .12366176
   _Iyear_87 | -1.3247366  .87641421 -1.5115416  .06532525
   _Iyear_88 | -.93795347  1.0381108 -.90351959   .1831251
       y-lag |  1.5332567  .06307817  24.307248          0
```

First the iteration logs are reported, then the estimation output is displayed in a standard fashion, where the first columns reports the estimated coefficients for the QE_basic model, along with asymptotic standard errors, the related t-statistics and p-values. Notice that the estimate associated with ψ in Equation (5) reflects a high degree of positive state dependence, in line with the well-known results in other applied works.

The extended version of the QE model, QE_ext, can be fitted in a similar manner, by using command `cquadext`:

```
cquadext union idcode age grade south not_smsa _Iyear_72 _Iyear_73
_Iyear_77 _Iyear_78 _Iyear_80 _Iyear_82 _Iyear_83 _Iyear_85 _Iyear_87
```

Notice that here we are not using the `xi:` prefix and the factor `i.year` as explanatory variable. In fact, we type in the time dummies separately in order to exclude the dummy for 1988: in the QE_ext model, not all the effects associated with the time dummies can be identified, because of the presence of an intercept term, ϕ, in the covariates referred to the observation at time T (see Equation (4)).

The above code produces the following output:

```
. cquadext union idcode age grade south not_smsa _Iyear_72 _Iyear_73
> 2 _Iyear_77 _Iyear_78 _Iyear_80 _Iyear_8 _Iyear_83 _Iyear_85 _Iyear_87

Fit quadratic exponential model by Conditional Maximum Likelihood
see Bartolucci & Nigro (2010), Econometrica
```

output omitted

```
               |      est.      s.e.      t-stat.    p-value
---------------+-----------------------------------------------------
           age |  .17308473  .11933765  1.4503782  .07347655
         grade | -.04047509   .0465145 -.87016079  .19210627
         south | -.51184847  .13953697 -3.6681926  .00012214
      not_smsa |  .17524652  .13523937  1.2958248  .09751793
     _Iyear_72 |  -.4644361   .1964388 -2.3642789   .0090326
     _Iyear_73 | -.65950516  .27895047 -2.3642375  .00903361
     _Iyear_77 | -1.3784265  .72358421 -1.9049981  .02839016
     _Iyear_78 | -1.3701126  .84614133 -1.6192479  .05269697
     _Iyear_80 | -1.1167485  1.0780889 -1.0358595  .15013386
     _Iyear_82 | -1.9383478  1.3150617 -1.4739595  .07024624
     _Iyear_83 | -2.4862166   1.433189 -1.7347444  .04139305
     _Iyear_85 |  -2.293721  1.6709237 -1.3727264  .08491871
     _Iyear_87 | -2.8867738  1.9100228 -1.5113819  .06534559
      diff-int | -2.9745408  2.2316307 -1.3329001   .0912823
      diff-age |  .01050808  .02053247  .51177844  .30440304
    diff-grade |  .01403913  .02483142  .56537754   .2859085
    diff-south | -.01017179  .12702618 -.08007635  .46808827
```

```
 diff-not_smsa | -.24502608  .14435482 -1.6973876   .0448117
diff-_Iyear_72 |  4.5353507  2.3909244  1.8969026   .0289204
diff-_Iyear_73 |   3.213293  2.1675763  1.4824359  .06911217
diff-_Iyear_77 |  2.9858792  2.1187489  1.4092653  .07937837
diff-_Iyear_78 |  2.8557536  2.1441469  1.3318834  .09144926
diff-_Iyear_80 |  3.4787453  2.1165322  1.6436061   .0501288
diff-_Iyear_82 |  2.3113123   2.108686  1.0960913  .13651941
diff-_Iyear_83 |  2.4132524  2.1023472  1.1478848  .12550806
diff-_Iyear_85 |  2.7441521  2.0885294   1.313916  .09443724
diff-_Iyear_87 |  2.6838554   2.079152  1.2908414  .09837934
         y-lag |  1.5646588  .06439017  24.299654          0
```

where the iteration logs have been omitted for brevity. If the time-dummy associated with the last observation is not dropped beforehand, a warning message is printed, and results are obtained using the generalized inverse of the Hessian.

The modified QE model, QE_equ, can be estimated by calling `cquadequ`

```
xi: cquadequ union idcode age grade south not_smsa i.year
```

```
. xi: cquadequ union idcode age grade south not_smsa i.year
i.year            _Iyear_70-88        (naturally coded; _Iyear_70 omitted)
```

output omitted

```
             |      est.      s.e       t-stat.    p-value
-------------+----------------------------------------------
         age |  .16845566  .11901965  1.4153601  .07848147
       grade | -.03958659  .04550678 -.86990548  .19217603
       south | -.53406297  .13625918  -3.919464  .00004437
    not_smsa |   .0984639  .13080979  .75272577  .22580736
   _Iyear_71 |  1.6032853  1.0337023  1.5510126  .06044933
   _Iyear_72 |  1.1740137  .91650676  1.2809657  .10010286
   _Iyear_73 |  .97015581  .79589985  1.2189421  .11143309
   _Iyear_77 |  .24177005  .33043231  .73167798  .23218257
   _Iyear_78 |  .25282926  .21264697  1.1889624  .11722723
   _Iyear_80 |  .54363568  .08483378  6.4082453  7.360e-11
   _Iyear_82 |  -.3246461   .2861711 -1.1344475  .12830343
   _Iyear_83 | -.88650878  .40228033  -2.203709  .01377241
   _Iyear_85 | -.68779397  .63653421 -1.0805295  .13995324
   _Iyear_87 | -1.3316314  .87497451 -1.5219087  .06401597
```

```
  _Iyear_88 | -1.5551096  1.0362781 -1.5006681  .06672071
      y-lag |  .76891417  .03180295  24.177448          0
```

The estimation results are marginally different from those obtained by `cquadbasic` because of the different association between y_{it} and y_{it-1} in Equation (6), used to exploit more information in testing for state dependence. The test for absence of state dependence is the t-test associated with the lagged dependent variable reported in the output above.

Finally, command `cquadpseudo` fits the pseudo-CML estimator of the parameters of the DL model described in Section 2.4. The input line is as follows

```
 xi: cquadpseudo union idcode age grade south not_smsa i.year
```

and produces the following output

```
. xi: cquadpseudo union idcode age grade south not_smsa i.year
i.year            _Iyear_70-88        (naturally coded; _Iyear_70 omitted)

Fit Pseudo Conditional Maximum Likelihood estimator for the dynamic logit model
see Bartolucci & Nigro (2012), J.Econometrics

First step

      |      lk       lk-lk0
------+--------------------------------------------------
    1 | -4550.1859  1.000e+10
    2 | -4508.4587  41.727174
    3 | -4479.6267  28.832058
    4 | -4464.3395  15.287228
    5 | -4462.0772  2.2622144
    6 |  -4462.077   .0002431
    7 |  -4462.077  1.000e-11

Second step

      |      lk       lk-lk0
------+--------------------------------------------------
    1 | -3386.3831  1.000e+10
    2 | -3072.2352  314.14795
    3 | -3068.2783  3.9568752
```

```
    4 | -3068.2768  .00144833
    5 | -3068.2768  5.689e-10


             |      est.     s.e.(rob)    t-stat.    p-value
-------------+----------------------------------------------------
         age |  .18590097  .12502643  1.4868934  .13704297
       grade | -.03115066  .05488738 -.56753782  .57034884
       south | -.62116171  .16083689 -3.8620598  .00011244
    not_smsa |  .10764683  .14923884  .72130574  .47072142
   _Iyear_71 |  .66895192  1.0824925  .61797374  .53659265
   _Iyear_72 |  .26741545  .96467342  .27720827  .78162019
   _Iyear_73 |  .04473093  .83125482  .05381134  .95708548
   _Iyear_77 | -.66439033   .3474518 -1.9121798  .05585313
   _Iyear_78 | -.56283602  .22525051 -2.4987114  .01246458
   _Iyear_80 | -.42448135  .08815153 -4.8153602  1.469e-06
   _Iyear_82 | -1.3962766  .30058041 -4.6452681  3.396e-06
   _Iyear_83 | -1.8777382  .42388142 -4.4298667  9.429e-06
   _Iyear_85 | -1.7545693     .66529    -2.6373  .00835689
   _Iyear_87 |  -2.409943  .91783499 -2.6256822  .00864755
   _Iyear_88 | -2.5102739  1.0890873 -2.3049337  .02117029
       y-lag |  1.6295114  .07720721  21.105691          0
```

The first part of the output reports the value of the log-likelihood at each iteration for the first step, the CML estimation of the regression coefficients using a static logit model, while the second refers to the maximization of the pseudo log-likelihood with respect to the parameters in Equation (8). The estimation results are similar to those obtained with the QE model, nevertheless these results are a closer approximation of the true parameters of the DL model.

5 Acknowledgments

We acknowledge the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project "Mixture and latent variable models for causal inference and analysis of socio-economic data").

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, B*, 32:283–301.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. John Wiley & Sons.

Bartolucci, F., Bellio, R., Sartori, N., and Salvan, A. (2014). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews*, in press.

Bartolucci, F. and Nigro, V. (2010). A dynamic model for binary panel data with unobserved heterogeneity admitting a $\sqrt{n}$-consistent conditional estimator. *Econometrica*, 78(2):719–733.

Bartolucci, F. and Nigro, V. (2012). Pseudo conditional maximum likelihood estimation of the dynamic logit model for binary panel data. *Journal of Econometrics*, 170(1):102–116.

Bartolucci, F., Nigro, V., and Pigini, C. (2015). Testing for state dependence in binary panel data with individual covariates. *Econometric Reviews*, in press.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *The Review of Economic Studies*, pages 225–238.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In Heckman, J. J. and Singer, B., editors, *Longitudinal analysis of labor market data*. Cambridge University Press: Cambridge.

Cox, D. (1972). The analysis of multivariate binary data. *Applied statistics*, 21:113–120.

Croissant, Y. and Millo, G. (2008). Panel data econometrics in r: The plm package. *Journal of Statistical Software*, 27(2):1–43.

Heckman, J. J. (1981a). Heterogeneity and state dependence. *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press: Cambridge MA. Manski CF, McFadden (eds).

Heckman, J. J. (1981b). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discreted data stochastic process. *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press: Cambridge MA. Manski CF, McFadden (eds).

Honoré, B. E. and Kyriazidou, E. (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica*, 68:839–874.

Hsiao, C. (2005). *Analysis of Panel Data*. Cambridge University Press, New York, 2nd edition.

Lucchetti, R. and Pigini, C. (2015). `DPB`:dynamic panel binary data models in `gretl`. Technical report, `gretl` working paper.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-122.

Rabe-Hesketh, S. (1999). GLLAMM: Stata program to fit generalised linear latent and mixed models. Statistical Software Components, Boston College Department of Economics.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2):301–323.

Stewart, M. (2006). Maximum simulated likelihood estimation of random-effects dynamic probit models with autocorrelated errors. *Stata Journal*, 6(2):256–272.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics*, 20(1):39–54.