NONPARAMETRIC REGRESSION ON PANEL DATA MODELS WITH BINARY RESPONSES: A SURVEY

Ke Yang, Univesity of Hartford

ABSTRACT

This paper serves as an instroduction for economists to the field of nonparametric regression models for panel data with binary responses. The focus is on how dependences across observations present challenges to estimating the mean function efficiently with nonparametric methods. The purpose of this paper is to examine the basic motivation of different procedures, cover some theoretical results and bandwidth selection, discuss the relative performances of different methods, e.g. the nonparametric quasi-likelihood estimation method and other more recently developed methods targeting on improving the estimation accuracy using panel data structure. **JEL Classification:** C01, C14

INTRODUCTION

Panel data with a large number of individuals, with each individual measured only over a short period of time, is very common in economic studies. Parametric models have long been used to study the behaviorial relationships for such data (see Woodrige, 2000 for a good survey on the parametric approach). Very often, one would like to relax our assumption about the parametric functional form that captures the relationship between the dependent and independent variables (see pages 266-169 in Pagon and Ullah, 1999 and chapter 8 in Li and Racine, 2007 for a survey on nonparametric regression models). As a consequence there has been a growing literature in estimating the conditional mean function nonparametrically, see Frolich (2006), Hoderlein et al. (2011), Liang (2012) for some examples. Many competing nonparametric methods are currently available, including kernel-based methods, regression splines, smoothing splines and wavelet and Fourier series expansions. Among them, the local linear estimator has been accepted as an attractive nonparametric method of estimating the regression function and its derivatives, due to its appealing properties including its simplicity in computation, better boundary performance and minmax property (Fan, 1992). However, since the standard local linear estimator does not utilize the information in the error structure in its weighing scheme, more recently, many researches have focused on how to make the most efficient use of the panel data structure to improve the accuracy of the estimating behavioral relationship.

As it turns out, the problem is far from simple. An obvious method is to incorporate the covariance matrix of the error through weighted quasi-likelihood function (Severini and Staniswalis, 1994). However, this method produces an estimator with far worse asymptotic variance, even compared to simple local linear regression on the pooled data. An alternative method - "component estimator" that fits separate regression models on each component (panel) in the data and then combines these estimators to produce an overall estimator - though more complicated to implement, does not utilize the panel data structure and is asymptotically equivalent to the pooled estimator.

More recently, a few other procedures have been proposed to improve the local linear estimation by taking into consideration the information in the non-spherical error. Ruckstuhl et al. (2000) propose a revised version of quasi-likelihood estimator, which effectively ignores the with-in panel correlation completely and treat the data as if they are independent. Lin and Carroll (2000) refer to it "working independence" method. They show that this working independence estimator has a better asymptotic properties compared to quasi-likelihood estimator.

This result is "counter intuitive" (Wang, 2003) in that the nonparametric regression is actually better-off by ignoring some information in the data. As part of the efforts to make proper use of the within panel correlation in local linear like estimator, a few other methods were proposed in recent literature. Ruckstuhl et al. (2000) propose a two step pre-whitening procedure which involves a correction of the residuals with the inverse of the covariance matrix. They show that this estimator could have a smaller variance compared to working independence estimator. However, the two-step estimation results in a more complicated bias expression that could be bigger or smaller than pooled estimator. Martins-Filho and Yao (2009) propose an alternative version of the pre-whitening estimator which implements an over-smoothing and adjustment of the scale of the correction in the first step. This alternative procedure does have a better asymptotic variance while maintain the same bias expression. While more appealing in asymptotic, how this over-smoothing helps the estimator's finite sample performance need further exploration.

Wang (2003) proposes another two step procedure to account for the with-in panel correlation in a different way. In this procedure, whenever an observation, say j'th, is identified as local, all the points in the same panel are used in the local averaging. To avoid the higher bias, the points in that panel other than j are used only through the residuals which are calculated from a previous regression. This estimator, like pre-whitening estimator, has a smaller variance but a much more complicated bias expression compared to working independence estimator.

For empirical researchers, how to choose among alternative nonparametric procedures and apply it to answer an economic question is of essential interest. First, it is well known that non-parametric estimators, in addition to be computational intensive, also have slower convergence rates compared with parametric estimators, which can present some challenges given that empirical researchers always face a finite sample. Furthermore performances of nonparametric estimators rely largely on the choice of a smoothing parameter, the bandwidths. In practice bandwidths can be chosen arbitrarily by the researchers or using some data driven methods, such as cross validation, and various plug-in methods (see for example Mammen and Park, 2005). Therefore, a carefully selected survey on these available estimators with discussion on data driven bandwidth selection methods would be valuable to emperical researchers who are interested in applying these procedures in various scenarios.

This paper explores how ach of these methods are contructed to improve both the robustness and the accuracy of the stimation. To trace the evolution of nonparametric methods as a natural extension of their parametric analogy, a brief review on some parametric approaches, both likelihood-based and method of moment based, is provided. Regarding the nonparametric methods, the focus is on how each method utilizes the dependence structure of the data and how they compare with each other in their estimation accuracy. In addition some current and potential research topics in this area are discussed. Other than this section, the paper is organized as follows. In section 2, the logit panel model under study is defined and two well-known parametric methods, i.e. maximum likelihood estimation and generalized estimating equation method, are reviewed. In section 3, the nonparametric quasi-likelihood estimator and a revised version of it, "working independence" estimator are discussed. In section 4, some more recently proposed methods are discussed and compared to the quasi-likelihood estimator. In section 5, the paper concludes with a summary on the key features of each estimator and a discussion on some potential research topics in the area.

PANEL DATA WITH BINARY RESPONSES: PARAMETRIC METHODS

Consider a nonparametric panel data model in which it is assumed that there are n individuals in the sample and each individual is measured at the same J occasions. The response variables are binary (0 or 1) and the covariates could be either discrete or continuous variables. Let $\vec{Y}^i = (y_{i1}, \cdots, y_{iJ})'$, where y_{ij} is the binary random variable and takes value 1 if individual i has response 1 (success) at occasion j, and 0 otherwise. Similarly, let $\vec{X}^i = (x_{i1}, \cdots, x_{iJ})'$, a $J \times p$ where x_{ij} is a $p \times 1$ vector representing the covariate vector for individual i at occasion j. The whole sample points is represented by a $(n \times J)$ by $p+1$ vector, $(\vec{Y}, \vec{X})$, where $\vec{Y} = (y_{i1}, \cdots, y_{nJ})'$ and $\vec{X} = (x_{i1}, \cdots, x_{nJ})'$. The marginal distribution of is assumed to be Bernoulli,

$$f(y_{ij} \mid \vec{X}^i) = exp[y_{ij}\theta_{ij} - log(1 + exp(\theta_{ij}))]. \tag{1}$$

In a logit regression we assume

$$\begin{aligned} \theta_{ij} &= log[\mu_{ij}/(1 - \mu_{ij})] = x_{ij}'\beta \\ \mu_{ij} &= \mu_{ij}(\beta) = E(Y_{ij}) = prob(Y_{ij} = 1 \mid x_{ij}) \end{aligned} \tag{2}$$

where is the probability of success at occasion j; $v_{ij}(\mu_{ij}) = Var(Y_{ij} \mid \vec{X}^i) = \mu_{ij}(1 - \mu_{ij})$ is the variance function for Y_{ij}; and β is a $p \times 1$ vector of parameters that are of our interests to estimate. Let $\mu_{\bar{ij}}(\beta) = E(\vec{Y}^i) = (\mu_{i1}, \cdots, \mu_{iJ})'$ be the vector containing the marginal probabilities of success for individual at all occasions.

This is a typical set up for a parametric logit regression. Note that up to now, we have not specified a within-subject correlation structure for the individual i yet.

Likelihood-Based Method

In the case where responses are independent, a regression method can be derived based on the joint distribution of the binary responses,

$$f(\vec{Y} \mid \vec{X}) = \prod_{i=1}^{n} exp\left(\sum_{j=1}^{J} y_{ij}\theta_{ij} - \sum_{j=1}^{J} log(1+exp(\theta_{ij})) \right). \tag{3}$$

which leads to the log-likelihood function taking the following form,

$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[y_{ij}\theta_{ij} - log(1+exp(t \right. \tag{4}$$

Then by taking derivative with respective to β of the likelihood function we have

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n}\left(\frac{\partial \vec{\mu}^{i}}{\partial \beta}\right)' \frac{\partial \vec{\theta}^{i}}{\vec{\mu}^{i}} \frac{\partial L_i}{\partial \vec{\theta}^{i}} \tag{5}$$

The exponential family distributions have the properties that the derivative of the log-likelihood with respect to the canonical parameter, $\vec{\theta}^{i}$, is

$$\frac{\partial L_i}{\partial \vec{\theta}^{i}} = \vec{Y}^{i} - \vec{\mu}^{i} \quad \text{and} \quad \frac{\partial \vec{\mu}^{i}}{\partial \vec{\theta}^{i}} = Cov(\vec{Y}^{i})^{i} \tag{6}$$

and additionally, by assumption, $\mu_{ij} = exp(x_{ij}'\beta)/[1+exp(x_{ij}'\beta)]$, so we have

$$\left(\partial \vec{\mu}^{i} / \partial \beta\right) = (\vec{X}^{i})'\Delta_i \tag{7}$$

where $\Delta_i = diag\{Var(Y_{i1}),\cdots,Var(Y_{iJ})\}$. Hence, the derivative of the likelihood function with respect to the parameters vector β can be written as

$$\partial L / \partial \beta = \sum_{i=1}^{n} \partial L_i \partial \beta = \sum_{i=1}^{n} (\vec{X}^{i})'(\vec{Y} - \vec{\mu}_i) = 0 \tag{8}$$

where μ_{ij} is an function of $\hat{\beta}$, the estimate of β as defined in equation (2). The maximum likelihood estimator of β is the solution to above nonlinear equation system.

Liang and Zeger (1986) show that even though the estimator defined in equation (8) ignored the correlation structure among the measurements for the same individual, it remains to be consistent and asymptotically normal. However, the inverse of the estimated information matrix yields an inconsistent estimator for the variance parameters. Liang and Zeger (1986) propose a "robust" variance estimator that is an analogue of the "White robust covariance estimator" in the linear model. This estimator is consistent regardless the true dependence structure among responses.

Generalized Estimating Equation (GEE)

The moment-based GEE approach in Liang and Zeger (1986) produces consistent estimator of β, given the mean function, μ_{ij}, is specified correctly. The GEE estimator for β are the solution to the following equation system,

$$\sum_{i=1}^{n}(D^i)'(V^i)^{-1}(\vec{Y}^i - \vec{\mu}^i) = 0 \tag{9}$$

where $D^i = \partial\vec{\mu}^i/\partial\beta'$ and V^i is a "working" covariance matrix of $\vec{Y}^i$ chosen arbitrarily by the researcher. The "working" covariance is defined to take the form $V^i = \Delta_i^{1/2}R_i(\alpha)\Delta_i^{1/2}$ where $\Delta_i = diag\{Var(Y_{i1}),\cdots,Var(Y_{iJ})\}$, and $R_i(\alpha) = Corr(\vec{Y}^i)$ is a $J \times J$ "Working" correlation matrix defined through the vector of parameters α. The equation system defining $\hat{\beta}$ is of the same form to the quasi-likelihood estimating equations introduced in McCullagh and Nelder (1989, Ch.9). Comparing the estimator defined in equation (8) and the one defined in (9), equation (9) is a generalization of (8) by introducing the "working" correlation matrix, $R_k(\alpha)$. By plugging in the functional form for $(D_i)'$ given by the binary model specification in (1), we obtain the following estimating equation

$$\sum_{i=1}^{n}(\vec{X}^i)'\Delta_i V_i^{-1}(\vec{Y}^i - \vec{\mu}^i) = 0 \tag{10}$$

Some common specifications for the "working" correlation matrix $R_i(\alpha)$ are as follows:

1. $R_i(\alpha) = I_J$, a $J \times J$ identity matrix. This corresponds to the "working independence" assumption that essentially ignores the within subject correlation and gives estimating equations identical to equation (8). No alpha needs to be estimated.

2. Constant correlation, i.e. $R_i(\alpha) = (1-\alpha)I_J + \alpha\vec{1}_J\vec{1}_{J'}$ where I_J is an $J \times J$ identity matrix and $\vec{1}_J$ is a $J \times 1$ one vector, α is a scaler correlation coefficient remained to be constant across all measurements for the same individual and needs to be estimated.
3. Autoregressive correlation, i.e. $Corr(Y_{is}, Y_{it}) = \alpha^{|s-t|}$. This specification is common when there exists a natural ordering of the

measurements in time.

4. Unstructured Correlation: $Corr(Y_{is}, Y_{it}) = \alpha_{st}$, In this case, $R_i(\alpha)$ has $J(J-1)/2$ different elements to be estimated.

Many other specifications of the correlation structure are available. One of the most attractive features of the GEE approach is that it provides a consistent estimate of the regression parameter β that only requires the correct specification of the mean function, regardless of whether the $R_i(\alpha)$ is correctly specified.

QUASI-LIKELIHOOD ESTIMATION: NONPARAMETRIC METHODS

Likelihood-Based Local Linear Regression For Independent Data

One can extend the logit model (1) to a nonparametric model where θ_{ij} is assumed to to be a smooth but otherwise unspecified function. Consider a local polynomial of order one regression, i.e. local linear regression, (Fan, 1992), with a bandwidth h, and the symmetric kernel density function is $K(\cdot)$, normalized without loss of generality to have unit variance. Define $K_h(v) = 1/h \cdot K(v/h)$. The basic idea for local linear regression is to approximate $\theta(\cdot)$ at a given point x_0 using a first order Taylor expansion, $\vec{\theta}^i(\vec{X}^i) \approx \left(\vec{G}^i(x_0)\right)\beta$, where $\vec{G}^i(x_0) = \{\vec{1}_J, \vec{X}^i - \vec{1}_J x_{0'}\}$ and β is a $(p + 1)$ x 1 vector with the first element being $\theta(x_0)$, the mean function evaluated at x_0, the value of which is of our interests in estimation.

If one naively assumes that the responses are all independent, then the likelihood function of the binary responses weighted by the kernel function can be written as,

(11)
$$L = \sum_{i=1}^{n} L_i = \sum_{i=1}^{n}\sum_{j=1}^{n}\left[y_{ij}\theta_{ij} - log(1 + exp(\theta_{ij}))\right]K_h(x_{ij} - x_0).$$

Using the maximum-likelihood (ML) procedure discussed in section 2.1, and replacing the function $\vec{\theta}^i(\vec{X}^i)$ with its local approximation $\left(\vec{G}^i(x_0)\right)\beta$, we obtain the local linear ML estimating equations as follows,

$$\partial L / \partial\beta = \sum_{i=1}^{n} \partial L_i / \partial\beta = \sum_{i=1}^{n}(\vec{G}^i(x_0))'\vec{K}_h^i(x_0)(\vec{Y} - \vec{\mu}_i(x_0)) = 0. \qquad (12)$$

where $\vec{K}_h^i(x_0) = diag\{K_h(x_{i1} - x_0), \cdots, K_h(x_{iJ} - x_0)\}$, and $\vec{\mu}_i(x_0) = \vec{\mu}(\ \vec{G}^i(x_0))'\beta)$ is the local linear approximation (at x_0) of the mean function evaluaated at sample points. The local linear estimator, $\hat{\theta}(x_0, h)$ is the first element of the vector, $\beta = (\beta_0, \beta_1, \cdots, \beta_p)$ that solves the above nonlinear equation system.

Quasi-likelihood Estimator and Working independence Estimator

Severini and Staniswalis (1994) propose to incorporate the covariance matrix Σ in the construction of the nonparametric estimator, using an approach called quasi-likelihood maximizing. Lin and Carroll (2000) investigate this idea and show that this quasi-likelihood estimator is of the same form as the nonparametric extension of parametric GEE (10). The local linear version of this quasi-likelihood estimator, is defined to be the first element of the solution vector to either one of the following equation:

$$\sum_{i=1}^{n} (\vec{G}^i(x_0))' \Delta_i(x_0) V_i(x_0)^{-1} \vec{K}_h^i(x_0)(\vec{Y} - \vec{\mu}_i(x_0)) = 0 \tag{13}$$

or

$$\sum_{i=1}^{n} (\vec{G}^i(x_0))' \Delta_i(x_0) (\vec{K}_h^i(x_0))^{1/2} V_i(x_0)^{-1} (\vec{K}_h^i(x_0))^{1/2} (\vec{Y} - \vec{\mu}_i(x_0)) = 0 \tag{14}$$

where $\{\vec{\mu}^i(x_0), \Delta_i(x_0), V_i(x_0)\}$ are evaluated at $\vec{\mu}_i(x_0) = \vec{\mu}(\, G^i(x_0))'\beta)$. The Fisher scoring algorithm can be used to solve the above two equations to get an estimator for β. In practice, the estimator $\hat{\beta}$ can be updated using iteratively re-weighted least square to solve

$$\sum_{i=1}^{n} (\vec{G}^i(x_0))' C_i(x_0) \vec{G}^i(x_0) \beta = \sum_{i=1}^{n} (\vec{G}^i(x_0))' C_i(x_0) \vec{Z}^i \tag{15}$$

where $C_i(x_0) = \Delta_i(x_0) V_i^{-1} \vec{K}^i(x_0) \Delta_i(x_0)$ is a working weight matrix and $\vec{Z}^i = \vec{G}^i(x_0)'\beta + \Delta_i^{-1}\{\vec{Y}^i - \vec{\mu}^i(x_0)\}$ is a working vector.

Asymptotic properties of the $\hat{\theta}(x_0, h)$ under (13) is difficult to derive for a general working correlation matrix R_i and non-Gaussian data. In the special case of Gaussian data with $\Delta_j(\cdot) = 1$ and any "working" covariance matrix, Lin and Carroll (2000) show that $\hat{\theta}(x_0, h)$ has asymptotic bias equal to $1/2 \cdot h^2 \theta^{(2)}(x_0)$ and asymptotic variance equal to $c/(nh)$, where the expression of c is complicated and its expression can be found in Ruckstuhl et al. (2000, Appendix). Note that the asymptotic expression for the bias and variance is free of the distribution of $\vec{X}^i$. This result indicates that quasi-likelihood estimator is consistent and converges at a rate of $\sqrt{1/nh}$, regardless of the form of the "working" covariance matrix being used. In particular, when observations on the same individual are independently and identically distributed with marginal density $f(\cdot)$ and variance σ_X, the bias and variance of the quasi-likelihood estimator has an explicit expression as follows.

$$E[\hat{\theta}(x_0,h)-\theta(x_0)]=\frac{h^2}{2}\theta^{(2)}(x_0)+o(h^2) \tag{16}$$

and

$$V[\hat{\theta}(x_0,h)]=\frac{\gamma(0)}{nJh\cdot f(x_0)}\left\{1+C_J\left(\frac{f^{(1)}(x_0)}{f(x_0)}\right)^2\right\}^{-1}+o(1/nh) \tag{17}$$

where $d_J=J\rho/(1+(J-1)\rho)$ and $C_J=(\frac{d_J}{(J-d_J)})^2(J-1)\sigma_X^2$. In addition, the above variance is minimized when the "working" correlation matrix is taken to be $R_i=I_J$, and the variance is

$$Var(\hat{\theta}(x_0,h))=\frac{\gamma(0)}{nh}\left[f(x_0)\sum_{j=1}^{J}1/\sigma_{jj}\right]^{-1} \tag{18}$$

Essentially, this result indicates that in local linear regression, at least for Gaussian data, it is optimal to simply ignore the within-subject correlation. Whether this result holds for logit model as specified in (1) is still in question.

The approach of ignoring the within-subject corelation, known as "working independence" estimator is the most efficient estimator for model (1). To be precise, this "working independence" estimator can be defined as the first element of the solution vector of the following equation system,

$$\sum_{i=1}^{n}\vec{G}^i(x_0)'\Delta_i(x_0)[diagV_i(x_0)]^{-1}\vec{K}_h^i(x_0)(\vec{Y}-\vec{\mu}_i(x_0))=0. \tag{19}$$

Researchers accustomed to correcting standard errors for dependence usually believe that adjusting for the within-subject correlation should improve the estimation accuracy, as in parametric regression. Wang (2003) argues that this resullt is counter intuitive in that correctly specifying the within-subject correlation actually leads to an asymptotically less efficient estimator.

Pre-Whitening Estimator (PW)

Ruckstuhl et al. (2000) and Martins-filho and Yao (2009) propose a “pre-whitening” estimator for the Gaussian data, here we extend it to the model (1) as follows:

Step 1: Run a local linear working independence regression of $\vec{Y}$ on $\vec{X}$ and construct an arbitrary vector $\vec{Z}$, with ij’th element being z_{ij}, as follows:

$$\vec{Z} = \tau\hat{\Omega}^{-1/2}\vec{Y} - (\tau\hat{\Omega}^{-1/2} - I)\hat{\vec{\theta}}(\vec{X}, h_s) \tag{20}$$

where $\hat{\Omega}$ is a consistent estimator of Ω, the true covariance matrix of the response variable, τ is a arbitrary scaler controlling the scale of the correction, h_s is a staring bandwidth used with kernel function and $\hat{\vec{\theta}}(\vec{X})$ is a vector of $\theta(\cdot)$ local linear working independence estimators of at sample points.

Step 2: The pre-whitening estimator is defined to be the local linear regression of $\vec{Z}$ on $\vec{X}$, i.e. the first element in $\vec{\beta}$ that soloves the following estimating equation,

$$\sum_{i=1}^{n} \vec{G}^i(x_0)' \vec{K}_h^i(x_0)(\vec{Z}^i - \vec{\mu}^i(x_0)) = 0 \tag{21}$$

Note that equation (20) is equivalent to

$$\vec{Z} = \hat{\vec{\theta}}(\vec{X}) + \tau\hat{\Omega}^{-1/2}[\vec{Y} - \hat{\vec{\theta}}(\vec{X})]. \tag{22}$$

Hence, it is clear that $\vec{Z}$ is a sum of a vector of local linear working independence estimates and the residual vector corrected by $\tau\hat{\Omega}^{-1/2}$. In practice, the performance of this estimator depends on the performance of $\hat{\Omega}$ and $\hat{\theta}(\vec{X})$.

Assuming that one bandwidth is used throughout, the bias and variance of the “pre s:

$$E[\hat{\theta}_{PW}(x_0, h) - \theta(x_0)] = -\frac{h^2}{2}\tau\left(\upsilon_d\theta^{(2)}(x_0) + \upsilon_o \frac{\sum_{l=1}^{J} f_l(x_0) \sum_{k \neq l}^{J} E(\theta^{(2)}(X_{1k} \mid X_{1l} = x_0))}{\sum_{j=1}^{J} f_j(x_0)}\right) + o(h^2) \tag{23}$$

and

$$Var[\hat{\theta}_{PW}(x_0, h)] = \gamma(0)\tau^2(nJh \cdot s(x_0))^{-1} + o(1/nh) \tag{24}$$

where υ_d and υ_o are diagonal and off-diagonal elements in $\Sigma^{-1/2}$. More specifically, they can be written as functions of element in Ω as suggested in Martins-Filho and Yao (2009),

Martins-Filho and Yao (2009) propose another version of "pre-whitening" procedure for Gaussian data, denoted here by **PW2**, that can be shown to be asymptotically normal. The **PW2** differs from the original version **PW** in two aspects. First, it utilizes an over-smoothing bandwidth in the first step regression. Second, it normalizes the $\hat{\Omega}^{-1/2}$ with its diagonal element. In particular, for model (1), the **PW2** is defined as follows,

Step 1: Construct $\vec{Z}$ as:

$$\vec{Z} = 1\upsilon_d\hat{\Omega}^{-1/2}\vec{Y} - (1\ \upsilon_d\hat{\Omega}^{-1/2} - I)\hat{\hat{\theta}}(\vec{X}) \tag{25}$$

where $\hat{\hat{\theta}}(\vec{X})$ is obtained with a starting bandwidth, h_s, such that $h_s h \to 0$ as $n \to \infty$.

Step 2: A local linear working independence regression of $\vec{Z}$ on $\vec{X}$, as defined in (21).

Note that in a special case that within subject correlation is zero, **PW2** is exactly the same as local linear estimator. The **PW2** estimator has different bias and variance expressions from **PW** estimator:

$$E[\hat{\theta}_{PW2}(x_0,h) - \theta(x_0)] = \frac{h^2}{2}\theta^{(2)}(x_0) + o(h^2) \tag{26}$$

and

$$Var[\hat{m}_{PW2}(x_0,h)] = \frac{\gamma_0(K)}{nJh}\upsilon_d^{-2}1s(x_0) + o(1/nh). \tag{27}$$

Note that when within subject correlation is zero, $\upsilon_d^{-2} = \sigma^2$, so the variance in (27) is exactly the same as local linear working independence estimator. When within subject correlation is not zero, one can show $\upsilon_d^{-2} \le \sigma^2$, there **PW2** is asymptotically more efficient than LL.

Marginal Kernel Estimator (MK)

Wang (2003) proposes a marginal kernel estimator for panel data as an improvement upon working independence estimator. One appealing property of MK is that the variance of the estimator is optimized when the correct Ω is used and this variance is asymptotically smaller than the working independence estimator. The MK estimator is defined to be

$\widetilde{\theta}(x_0, h) = \widetilde{\beta}_0$, where $\widetilde{\beta} = \{\widetilde{\beta}_0, \widetilde{\beta}_1, \cdots, \widetilde{\beta}_p\}$ solves the following equation:

$$0 = n^{-1}\sum_{i=1}^{j}\sum_{j}^{J} K_h(x_j - x_0)\Delta_i(x_0)\begin{bmatrix} 0 & \cdots, & 1 & \cdots, & 0 \\ 0\cdot\vec{1}_J, & \cdots, & \frac{x_j - x_0}{h}, & \cdots, & 0\cdot\vec{1}_J \end{bmatrix}(V^i)^{-1}\times\times\left\{\vec{Y}^i - \begin{bmatrix} \hat{\theta}(x_j, h_s) \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1\frac{x_j - x_0}{h} \\ \vdots \\ \hat{\theta}(x_j, h_s) \end{bmatrix}\right\} \tag{28}$$

where $\hat{\theta}(x_{ij}, h_s)$ is working independence estimator of $\theta(x_{ij})$ using a staring bandwidth h_s. The following assumption is made about the starting bandwidth h_s,

Assumption: The starting bandwidth h_s satisfies that $h_s^3 = o(n^{-1/2})$ and that $(nh_s)^{-1} = o(n^{-1/2})$. The bias and variance of marginal kernel estimator for Gaussian data, assuming that the true value of covariance matrix, Ω, is used for calculating the V_i, are given by:

$$E[\widetilde{\theta}(x_0, h) - \theta(x_0)] = \frac{h^2}{2}\theta^{(2)}(x_0) - \frac{h_s^2}{2}1Js(x_0)\frac{\mu_o}{\mu_d}\sum_{j=1}^{J}\sum_{l\neq j}^{J} E\{\theta^{(2)}(X_l) \mid X_j = x_0\} + o(h^2 + h_s^2) \tag{29}$$

and

$$Var[\widetilde{\theta}(x_0, h)] = \gamma_0(K)nJh1\mu_d s(x_0) + o(1/nh) \tag{30}$$

where μ_d and μ_o represent the diagonal and off-diagonal element in Ω^{-1}.

Wang (2003) shows that $1/\mu_d$ is always less or equal to σ^2, so **MK** estimator has a smaller variance than local linear "working independence" estimator. Compared to **PW2**, the relative size of the two variances depends on the ratio of $1/\mu_d$ to υ_d^{-2}, which in turn depends on the specific correlation structure of the error terms. The more concise expression of the bias expression for **PW2** is a result of over-smoothing in the first stage. However, whether this over-smoothing technique can improve the estimator's finite sample performance is not clear.

SUMMARY

In this paper some nonparametric kernel approaches to estimate a logit model with longitudinal/panel data structure are surveyed. The parametric GEE methodology yields $\sqrt{n}$ consistent estimates of the regression function provided that the model for the mean has been correctly specified. The nonparametric extension of GEE methodology provides consistent estimates without requiring the specifing the mean function, $\theta(\cdot)$. This robustness comes with a the cost of slower converging speed, $\sqrt{1/nh}$. Both methods produce consistent estimates regardless how the correlation matrix are specified. The GEE methodology does improve efficiency over the naive maximum likelihood estimator based on independent data, by using the correctly specified V_i. The similar result does not necessarily hold as we compare nonparametric regression approaches. The quasi-likelihood estimator, which utilizes a working covariance matrix, V_i, is shown to obtain the minimum asymptotic variance when the "working independence" covariance matrix is used, i.e., the estimation procedure is more accurate by ignoring the correlation structure. The other two methodologies, pre-whitening and marginal kernel estimator, does improve upon the "working independence" estimator in term of the asymptotic variance by using the correctly specified covariance structure.

Asymptotic results are useful largely to the extent that they can serve as a guide to what we may find in finite samples that researchers deal with in empirical studies. There are a few reasons that the asymptotic results may not serve us well in this respects. First, the asymptotic results only compare the leading term in the bias and variance expression while in finite sample, the ignored terms of smaller order asymptotically may have a major impact on the estimator's performance. Second, while a nonparametric regression allows researchers to estimate the mean function without specifying a parametric shape, they still need to choose a value for the smoothing parameter, the "bandwidth". It is well known that the performance of kernel estimators are very sensitive to the bandwidth selection. Typically, a bandwidth is chosen through minimization of the some criterion function like average mean squared error, which is calculated based on the estimator's asymptotic bias and variance therefore varies across estimators. How the bandwidths selected this way affect performance of the estimator is apparently not clear by comparing only the large sample results.

Therefore, a well-designed Monte Carlo experiment will serve a useful purpose in increasing our understanding of the problem of nonparametric estimation of economic relationship from a panel data structure. To make a fair comparison, the optimal bandwidth for each estimator could be calculated based on minimizing their asymptotic approximation of mean squared errors. In simulation studies, the optimal bandwidth can be truly optimal in the sense that the true model specification if known to the researcher. Some particularly interesting aspects of panel data model could be used in the simulation study to investigate their the impact on the performance of the estimators. These include the impact of the relative scale of the noise, different levels of the within subject dependency, the number of individuals in the sample relative to the number of measurements for each individual,and so on. Also, one could investigate the differences in the performance of the estimators that can be attributed to estimation of some intermediate values like V_i. These results could provide applied researchers with information that allows for a better understanding of these competing estimation alternatives in finite sample settings.

REFERENCES

Fan, J., 1992, Design adaptive nonparametric regression, *Journal of the American Statistical Association*, 87, 998-1004

Fitzmaurice, G.M., Laird, N.M. and Rotnitzky, A.G. , 1993, Regression Models for Discrete Longitudinal Responses, *Statistical Science*, 8, 284-309

Frolich, M., 2006, Non-parametric regression for binary dependent variables, *The Econometrics Journal*, volume 9, pages 511-540.

Hoderlein, S., Mammen, E. and Yu, K., 2011, Non-parametric models in binary choice fixed effects panel data, *The Econometrics Journal*, Volume 14, Issue 3, Pages 351-367.

Liang, Z., 2011, Binary Response Correlated Random Coeffcient Panel Data Models, *Working paper*, Department of Economics, Texas A&M University, College Station, TX.

Lin, X. and R.J.Carroll, 2000, Nonparametric function estimation for clustered data when the predictor is measured without/with error, *The Journal of the American Statistical Association*, 95, 520-534.

Mammen, E. and Park, B.U., 2005, Bandwidth selection for smooth backfitting in additive models. *The Annals of Statistics*, 33, 1260–1294

Martins-Filho, C. and Yao, F, 2009, Nonparametric regression estimation with general parametric error covariance, *Journal of Multivariate Analysis*, 100, 309-333.

McCullogh, P. and Nelder, J.A. , 1989, Generalized Linear Models, 2nd Ed, Chapman and Hall, New York.

Pagan, A. and A. Ullah, 1999, Nonparametric Econometrics, Cambridge University Press, Cambridge, UK.

Ruckstuhl, A.F., A.H.Welsh and R.J.Carroll, 2000, Nonparametric function estimation of the relationship between two repeatedly measured variables, *Statistica Sinica*, 10, 51-71.

Severini, T.A. and J.G.Staniswalis, 1994, Quasi-likelihood estimation in semiparametric models, *The Journal of the American Statistical Association*, 89, 501-511.

Wang, N., 2003, Marginal nonparmetric kernel regression accounting for within subject correlation, *Biometrica*, 90, 43-52.

Woodridge, J. M., 2010, Econometric Analysis of Cross Section and Panel Data, 2nd Edtion, The MIT Press.