Journal of Economics and Administrative Sciences
2024; 30(142), pp. 449-464

P-ISSN 2518-5764
E-ISSN 2227-703X

Journal of Economics and Administrative Sciences (JEAS)

Available online at http://jeasiq.uobaghdad.edu.iq
DOI: https://doi.org/10.33095/kpscqv37

A Comparison Between Speckman And Bayesian Estimation Method Of A Semiparametric Balanced Longitudinal Data Model

Ahmed Abdulsamad Habeeb*
Department of Statistics
College of Administration & Economics,
University of Baghdad, Iraq
Ahmed.Thamer1101@ coadec.uobaghdad.edu.iq
***Corresponding author**

Qutaiba N. Nayef Al-Kazaz
Department of Statistics
College of Administration & Economics,
University of Baghdad, Iraq
dr.qutaiba @coadec.uobaghdad.edu.iq

Received:16/9/2023 Accepted: 6/11/2023 Published Online First: 30 /8/ 2024

Abstract:

This paper aims to use semi-parametric regression to balanced longitudinal data model, where the parametric regression models suffer from the problem of strict constraints, while non-parametric regression models, despite their flexibility, suffer from the problem of the curse of dimensionality. Consequently, semi-parametric regression is an ideal solution to get rid of the problems that parametric and non-parametric regression suffer from. The great advantage of this model is that it contains all the positive features included in the previous two models, such as containing strict restrictions in its parametric component, complete flexibility in its non-parametric component, and clarity of the interaction between its parametric and non-parametric components.

Based on the above, two methods were used to estimate a semi-parametric balanced longitudinal data model. The first is the Bayesian estimating method; the second is the Speckman method, which estimated the unknown nonparametric smoothing function by employing the kernel smoothing Nadaraya-Watson method. The Aim was to make a comparison between the Bayesian estimation method and the classical estimation method. Three different sample sizes were used in the simulation studies: 50, 100, and 200. The study results showed that the Bayesian estimating method is best at low variance levels (1,5), whereas the Speckman method is best at high variance level (10).

Paper type: A Research derived from Dissertation Ph.D.
Keywords: Bayesian, Speckman, Nadaraya &Watson, Semi-parametric, Balanced longitudinal data.

1. Introduction:

Longitudinal data has received great attention in studies because it takes into account the effect of time change as well as the effect of changing cross-sections to the observations. Longitudinal data can be defined as the data that can be obtained through repeated observations of a phenomenon around (n) cross-sections. Which may be countries, institutions, companies, cities, individuals... etc., and during a certain time (T). (Zeger et al, 2002)

In Longitudinal data the phenomenon under study changes on two levels, the change on the horizontal level represents cross-section data, while the change on the vertical level represents Time -time-series data.

Note that the term longitudinal data gives the same meaning as panel data in the literature of researchers, As for this research it uses the term longitudinal data. Longitudinal data has special models to represent it, which are special regression models for longitudinal data, which have relied for a long time on parametric regression models, which is the most common, However, it has been noted in some applied aspects that this type may not represent the phenomenon under study appropriately due to the behavior of some variables that may be parametric behavior and others non-parametric and does not take into account the nonlinear effects of explanatory variables on the response variable. Therefore, the resulting estimates for this type of regression can be misleading.

There is a second type is the non-parametric regression model that takes into account the nonlinear effect of variables, which are characterized by their high flexibility and depend on the smoothing of the data using weight functions, The most famous of which is the kernel function that is used to smooth the data. However, most researchers have noticed that non-parametric regression also suffers from the curse of dimensionality, Which occurs when the number of variables increases. The third type is semi-parametric regression, The great advantage of this model is that it contains all the positive features included in the previous two types and the clarity of the interaction between its parametric and non-parametric components. (Green, 2002 ; Su, 2011)

The emergence of the semi-parametric approach in regression models is the result of the complementarity between both parametric and non-parametric inputs because its idea comes from the additive models where the parametric and non-parametric components are combined in this model, Consequently, this type has gained wide acceptance in economic and social studies and other modern studies such as longitudinal data. The areas of use of the three types mentioned above (parametric, non-parametric, and semi-parametric) have been expanded to be applied in Bayesian theory. Bayes' theory of statistical inference depends on employing prior information about unknown parameters and considering these parameters as random variables. Hence, this information can be formulated as a probability distribution. It is called the prior probability density function (Prior p.d.f.), and this information is obtained from previous data and experiments or the theory that governs this phenomenon. Bayes' theorem also relies on the current sample information represented by the likelihood function of the observations.

1.1 literature review:

Below are some previous studies in the field of Classical and Bayesian estimation For example:

Abd-Alhafez and Rashid (2013) compared the classical cubic spline method and the cubic spline technique with robust M estimators with were to avoid the problem of pollution in the data and to estimate time-varying parameter functions for balanced longitudinal data. The robust M estimators proved their efficiency in the study.

Wang (2014) developed a proposed method that was discussed by each of the researchers (Henderson and Ullah) in their research in (2005) by finding estimates for the non-parametric random effects model for the Longitudinal data, through the two-step method. Through the analysis of simulation experiments, the efficiency of the proposed estimator for all prepared samples was reached, using the comparison standard (MSE),

Sadiq (2015) found the best estimators for the parameters of the Longitudinal data models in the presence of the drop-out problem, by using multiple models and different methods of estimators. She used the estimation methods (ML, REML, GEE, WGEE, MI-GEE) in addition to the method proposed, that the method proposed by the researcher outperformed them for all sample sizes, as well as with different drop-out of levels.

Abdul razzaq (2015) discussed the problem of missing data in the dependent variable Y of the longitudinal data model. In his research, he touched on several methods of estimation to harness them in estimating parameters free from the overlapping effect between each of Y and μ, including the Bayesian Multiple Imputation, and the (EM) algorithm, the (ECM) algorithm, with two methods (normal and focused), the (ECME) algorithm, with two methods (two methods of segmentation of parameters),

Khalil and Fadam (2016) studied the Mixed-effects conditional logistic regression in longitudinal pollution data. The research demonstrated that conditional logistic regression is a robust evaluation method for environmental studies. It was shown through simulation that mixed-effects conditional logistic regression is more accurate for pollution studies.

Shaker (2016) submitted a dissertation in which she used parametric and semi-parametric Bayesian methods to estimate the reliability of the systems using the Dirichlet process prior and compared them with the reliability estimations of the systems using the classical methods. The results showed the preference of the Bayesian method for a sample size of n=14.

Liu et al (2017**)** proposed methods for estimating the parameters of a non-parametric model for Longitudinal data, which was considered one of the important modeling options in the effect of the covariates variable that may change dynamically over time using the correlation function, The researchers also proposed a new method that includes the performance of the selected sample was evaluated by conducting simulation studies, and an example of real data was analyzed to illustrate the proposed methods.

Hamza (2018) studied the marginal slopes of the Longitudinal data model, where she considered them as random variables, where the random cross-sectional errors are characterized by heterogeneity of the variance as well as being correlated of the first degree sometimes (depending on the circumstances of the phenomenon) in both the random and fixed effects models. While the previous studies considered it fixed and used what is known as the swammy model, whose parameters are characterized by randomness.

Burhan and Hamoud (2018) compared the estimate of the transfer function using the non-parametric method, represented by two methods: positional linear regression, the cubic bootstrap method, and the semi-parametric method, represented by a single-indicator semi-parametric model with the proposed cubic bootstrap, and the study proved that the proposed estimator is the best among the studied estimators.

Abd-Alreda (2019) presented a master's dissertation in which he used the non-parametric and semi-parametric Bayesian methods to estimate the Cox regression model and the survival function using the Dirichlet process prior and compared it with the estimators of the survival function using the classical methods. The results showed that the best method was the semi-parametric Bayes method.

Castelein et al (2020) developed a general method for selecting heterogeneous variables in non-linear Longitudinal data models such as polynomial logarithms models based on the Bayesian semi-parametric method and Dirichlet process mixture (DPM) and they reached an improvement in performance in the process of selecting variables heterogeneous.

Kamel (2021) studied the non-parametric and semi-parametric estimators for the random effects model of Longitudinal data, using the non-parametric estimators (Nadaraya - Watson), and (Local Linear Polynomial), and the semi-parametric methods (Speckman) and (Profile Least Square).

Nayef and Lina (2022) estimated the missing values for the multivariate skew normal distribution function using the K-nearest neighbors Imputation (KNN). After estimating the missing values, the parameters are estimated using Genetic Algorithm (GA). and the Bayesian Approach was also used to estimate the missing values and find the estimates for the parameters. by comparing the two methods the (GA) that is based on the (KNN)algorithm to estimate the missing values proved to be better and more efficient than the Bayesian Approach in terms of the results.

Nayef and Ali (2022) estimated the analysis of stochastic differential equation with long memory, represented by fractional diffusion process, in this paper they suggested a method for a system of stochastic differential equations with long memory, Also they use the Bayesian methodology to incorporate the advanced knowledge, the proposed method has been proved to be very accurate.

1.2 The problem of research :

The problem of research is that parametric regression models suffer from the problem of strict constraints, while non-parametric regression models, despite their flexibility, suffer from the problem of the curse of dimensionality.

1.3 The research Aim :

The research Aim to Use Semi-parametric regression is an ideal solution to get rid of the problems that parametric and non-parametric regression suffer from, The great advantage of this model is that it contains all the positive features included in the previous two models, such as containing strict restrictions in its parametric component, complete flexibility in its non-parametric component.

2.Material and Methods:

2.1 Semi-parametric Regression:

Semi-parametric regression is defined as a model that contains two components, one of which is a parametric component that has finite dimensions and the other non-parametric has infinite dimensions. There are two approaches to parametric estimation, The first is to estimate the parameters component by any parametric estimation method in the first step then in the second step, the non-parametric part is estimated by any method of non-parametric estimation, depending on the estimates of the first step.

The second method is opposite to the first method, as the non-parametric component is estimated in the first step, and in the second, the parameter component is estimated based on the estimates of the first step.

In this paper, the method of Nadaraya-Watson mentioned in equation (3) was relied upon to estimate the non-parametric part, while the ordinary least squares method was used to estimate the parametric part.

2.2 Partially Linear Longitudinal data Model With Random Effect :

The Partially Linear Model (PLM) is one of the semi-parametric regression models. It is one of the models that depend on linear (parametric) variables and non-linear (non-parametric) variables, which are usually continuous. (Su, 2011) (Green, 2002), and it has several names, including the semi-parametric regression model (SPRM) or the partially linear model (PLM).

The semi-parametric partially linear model of longitudinal data is described by the equation below: (Yun,2012) (poo,2017)

$$y_{ij} = x'_{ij}\beta + m(z_{ij}) + \varphi_{ij}\,,\quad i = 1,\ldots,n\quad, j = 1,\ldots,t \tag{1}$$

Where y_{ij}, response variable vector with dimensional nt x 1, x_{ij} Matrix of parametric explanatory variables with dimensional nt x q, β The vector of the unknown parameters with dimensional p x1, $x'B$ The parametric part of the model under study, z_{it} It represents the nonparametric variable in the data with n x 1, $m(z_{it})$: The nonparametric part function is smoothing function with n x1, φ_{it} random error vector.

The semi-parametric linear partially model defined in formula (1) is subject to the following conditions and assumptions:

$$E(v_{ij}|x_{ij}, z_{ij}) = 0$$
$$E(v_{ij}v'_{ij}|x_{ij}, z_{ij}, u_i) = \sigma^2_{ij}I_t$$
$$E(u_i|x_{ij}, z_{ij}, u_i) = 0$$
$$E(u_i^2|x_{ij}, z_{ij}, u_i) = \sigma^2_u$$
$$E(\varphi_{ij}|z_{ij}) = 0$$

Therefore, the model in(1), according to the above assumptions, will be as follows:

$$E(y_{ij}|x_{ij}, z_{ij}) = x'_{ij}\beta + m(z_{ij}) \tag{2}$$

2.2.1 The Classical Estimation:

2.2.1.1 Nadaraya and Watson Estimation Method :

Nadaraya and Watson is considered of Non-parametric estimator with a widespread application in nonparametric regression, was proposed by Nadaraya and Watson in 1994, The two researchers developed a method based on the idea of a graph in estimation presented previously by (Tukey, 1961) called (Regressograms).

The Nadaraya and Watson estimation method for longitudinal data is expressed as follows: (Hardle, 2004)

$$\widehat{m}(x_{ij})_{NW} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{t} y_{ij}K(\frac{x_{ij}-x}{h})}{\sum_{i=1}^{n}\sum_{j=1}^{t} K(\frac{x_{ij}-x}{h})} \tag{3}$$

Where: $\widehat{m}(x_{ij})_{NW}$ Nadaraya-Watson estimator, k(.) Kernel smoothing function and the Gaussian function was used ,h bandwidth

2.3 Estimation Methods of the Model Semi-Parametric For Longitudinal Data with Random Effect:

In the following sections (2.3.1), We will review the semi-parametric methodsused in the research to estimate the semi-parametric model of the longitudinal data in equation (1) as:

2.3.1 Speckman Method :

This method was proposed in (1988) by (Speckman) It depends on the partial residuals to estimate the partially linear model. The steps of this method are summarized by estimating the non-parametric component by taking the conditional expectation of the model defined in equation (1) with respect to z_{ij} and as shown below: **(**Speckman, 1988 ; Poo, 2017)

$$E(Y_{ij}\backslash z_{ij}) = E(X_{ij}\backslash z_{ij})'\beta + E[m(z_{ij}\backslash z_{ij})] + E[m(\varphi_{ij}\backslash z_{ij})] \tag{4}$$

Under the conditions of the partially linear model, that $E[m(\varphi_{ij}\backslash z_{ij})] = 0$ and $E[m(z_{it}\backslash z_{it})] = m(z_{it})$, So by subtracting formula (4) from (1), we get:

$$Y_{ij} - E(Y_{ij}\backslash z_{ij}) = [X_{ij} - E(X_{ij}\backslash z_{ij})]'\beta + \varphi_{ij} \tag{5}$$

Where the non-parametric part is omitted from the partially linear model, and then this model is written in the form of matrices and vectors in the form below:

$$\hat{Y} = \hat{X}'\beta + \varphi_{ij} \quad (6)$$

When applying the method of least squares (OLS) to formula (6), estimates of the parameters β are obtained as follows:

$$\beta^* = (\hat{X}'\hat{X})^{-1}(\hat{X}'\hat{Y}) \quad (7)$$

Where:

$\hat{X}$: A matrix with dimensions (nt x q) whose rows represent $\hat{X}_{ij}$

$\hat{Y}$: An a row vector with dimensions (nt x 1) whose rows represent j-values

The two unknown terms can be estimated $E(Y_{ij} \backslash Z_{ij})$ or $E(X_{ij} \backslash Z_{ij})$ In formula (5) by using one of the non-parametric methods, let it be the method of Nadaraya -Watson, If we denote these two estimates $\hat{X}$ and $\hat{Y}$ respectively, then:

$$\hat{X} = \hat{E}(X_{ij} \backslash Z_{ij}) = \frac{1}{NTh^p} \frac{\sum_{r=1}^{N} \sum_{s=1}^{T} X_{rs} K_h(Z_{ij} - Z_{rs})}{\sum_{r=1}^{N} \sum_{s=1}^{T} K_h(Z_{ij} - Z_{rs})} \quad (8)$$

$$\hat{Y} = \hat{E}(Y_{ij} \backslash Z_{ij}) = \frac{1}{NTh^p} \frac{\sum_{r=1}^{N} \sum_{s=1}^{T} Y_{rs} K_h(Z_{ij} - Z_{rs})}{\sum_{r=1}^{N} \sum_{s=1}^{T} K_h(Z_{ij} - Z_{rs})} \quad (9)$$

By substituting formulas (8) and (9) into formula (7), we get an estimate of the parametric part of the partially linear model of the Longitudinal data:

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}(\hat{X}'\hat{Y}) \quad (10)$$

The nonparametric component of the partially linear model is estimated by minimizing the random errors of the first-order local linear polynomial estimation in the following formula:

$$\sum_{i=1}^{n} \sum_{t=1}^{t} [(Y_{ij} - X'_{ij}\hat{\beta}) - \gamma_0 - \gamma_1(Z_{ij} - Z)]^2 K_h(Z_{ij} - z) \quad (11)$$

To get solutions for both $\hat{\gamma}' = (\hat{\gamma}'_1 \, \hat{\gamma}'_0)$:

$$\hat{m}(Z) = \hat{\gamma} = I_{nt}(R'K_z R)^{-1} R' \, K_z(Y - X\hat{\beta}) \quad (12)$$

$$K_z(Y - X\hat{\beta}) = (Y_{11} - X_{ij}\hat{\beta}) \ldots (Y_{nt} - X_{ij}\hat{\beta})$$

R: matrix with dimension $(nt) \times (1 + p)$ where

$$R = \begin{bmatrix} 1 & \cdots & (z_{11} - z)' \\ \vdots & \ddots & \vdots \\ 1 & \cdots & (z_{nt} - z)' \end{bmatrix}$$

2.4 The Bayesian Estimation :

2.4.1 posterior distributions :

It is defined as a function that represents all the information about the parameters to be estimated after observing the sample information. It is also called distribution after sampling, The posterior distribution is denoted by the notation $\pi(\theta/D)$ and it is a conditional probability distribution for the parameter β with the condition that sample x is obtained and assuming that θ is a random variable that has a prior distribution and denoted by the notation $\pi(\theta)$, The inference of θ is based on the posterior distribution, which we obtain by Bayes' theorem, The posterior distribution of the random variable θ is given by the following formula:

$$\pi(\theta/D) = \frac{L(\theta/D) \, \pi(\theta)}{\int_\Theta L(\theta/D) \, \pi(\theta) d\theta} \quad (13)$$

Where: Θ represent the parameter range θ, $L(\theta/D)$ Likelihood function

From the above formula, it is clear that $\pi(\theta/D)$ is proportional to the Likelihood function multiplied by the prior distribution.

$\pi(\theta/D) \propto L(\theta/D)\,\pi(\theta)$

So it includes data contribution by $L(\theta/D)$ And the contribution of the primary information identified by $\pi(\theta)$.

2.4.2 Finding the estimation of the non-parametric component :

To achieve flexibility in applying Bayesian estimation, the nonparametric smoothing function m(z) in the model (1) will be transformed into a random effects component as in the semiparametric mixed models:(Mohaisen & Abdulhussein, 2014)

$y_{ij} = x'_{ij}\beta + m(z_{ij}) + \varphi_{ij}$

$z'_{ij}b_i = m(z_{ij})$

$y_{ij} = x'_{ij}\beta + z'_{ij}b_i + \varphi_{ij}$

$$y_{ij} - x'_{ij}\beta = z'_{ij}b_i + \varphi_{ij} \qquad (14)$$

Where:

z_{ij}: It is an n×q matrix of random explanatory variables

bi: A random effects vector with dimensions q×1

$$\varphi_{ij} \sim N(0, \tau^{-1}) \qquad (15)$$

Then the non-parametric model is estimated based on the nonparametric Dirichlet process.

2.4.2.1 Dirichlet Processes:

A Dirichlet process is a family of stochastic processes whose products are probability distributions. In other words, it is a probability distribution, over a set of probability distributions. It is often used in Bayesian inference to describe prior knowledge about the distribution of random variables, The Dirichlet process is specified by the base distribution, which is denoted by the base distribution (G_0) and a positive real number M called the scaling parameter (also known as concentration parameter). The base distribution is the expected value of the process that's means the Dirichlet process draws distributions "around" the base distribution in way like to the way a normal distribution draws real numbers around its mean. However, even if the base distribution is continuous, the distributions drawn from the Dirichlet process are almost surely discrete , The (M) specifies how strong this discretization is in the limit of M →0 , the realizations are all concentrated at a single value, while in the limit of M →∞ the realizations become continuous. Between the two extremes, The realizations are discrete distributions with less and less concentration as M increases.

It is worth noting that the Dirichlet process is used to find the prior distributions of non-parametric functions, Since then, it has been applied in the field of data mining, machine learning, arithmetic and counting, as well as in data science and information**. (Ferguson, 1973)**

To illustrate the Dirichlet process, It assumed that the random effects vector bi is distributed according to Dirichlet process where We assume G0 is a base distribution and M is a positive real number, Where we say that b is distributed according to Dirichlet operations with the base distribution G0 and the concentration parameter M and it can be written in the following formula: (Ferguson, 1973)

$$bi \sim DP(M, G_0) \qquad (16)$$

$$G_0 \sim N(0, D) \qquad (17)$$

Then the posterior conditional distribution of bi can be expressed according to the Dirichlet distribution on G^{-i} and G_0 as follows: (**Jochmann & León , 2004**).

Then the posterior conditional distribution of bi can be expressed according to the Dirichlet distribution on b^{-i} and G_0 as follows: (**Jochmann & León , 2004**).

$$b_i \backslash b^{-i}, G_0 \sim \frac{M}{M+N-1} G_0 + \frac{1}{M+N-1} \sum_{j=1}^{l} m_j^{-i} \delta(k_j^{-i}) \qquad (18)$$

P-ISSN 2518-5764
E-ISSN 2227-703X

Where:
G_0: base distribution
M: scale parameter
N: The distinct values of b
k^{-i}: cluster of distinct values of b^{-i}
m_j^{-i}: number of appearing times the distinct values of b^{-i}.
We can now merge this result with the Likelihood function that follows the normal distribution and take the integral we get: (**Jochmann & León, 2004**).

$$b_i\backslash\beta, D, G_0, b^{-i} \sim \int \emptyset\left(y_{ij}\backslash x'_{ij}\beta + z'_{ij}b_i, \tau^{-1}\right)d\left[b_i\backslash b^{-i}, G_0\right]$$

$$b_i\backslash\beta, D, G_0, b^{-i} \sim \int \emptyset\left(y_{ij}\backslash x'_{ij}\beta + z'_{ij}b_i, \tau^{-1}\right)\left[\frac{M}{M+N-1}G_0 + \frac{1}{M+N-1}\sum_{j=1}^{l} m_j^{-i}\delta\left(k_j^{-i}\right)\right] \quad (19)$$

We performing the integration we end up with:

$$b_i\backslash\beta, D, G_0, b^{-i} \sim \left[\frac{M}{M+N-1}\emptyset\left(y_{ij}\backslash x'_{ij}\beta + z'_{ij}b_i, \tau^{-1}\right) + \frac{1}{M+N-1}\sum_{j=1}^{l} m_j^{-i}\emptyset\left(y_{ij}\backslash x'_{ij}\beta + z'_{ij}b_i, \tau^{-1}\right)\right]$$

$$b_i\backslash\beta, D, G_0, b^{-i} \sim \frac{1}{M+N-1}\left[M\emptyset\left(y_{ij}\backslash x'_{ij}\beta + z'_{ij}b_i, \tau^{-1}\right) + \sum_{j=1}^{l} m_j^{-i}\emptyset\left(y_{ij}\backslash x'_{ij}\beta + z'_{ij}b_i, \tau^{-1}\right)\right] \quad (20)$$

The final posterior distribution of bi can be found as follows: **(west et al., 1994)**

$$p\left(bi\backslash\beta, \tau, y, b^{-i}\right) \propto \left\{M \int \emptyset\left(y_{ij}\backslash x_{ij}\beta + z_{ij}b_i, \tau^{-1}\right)\emptyset(b_i\backslash 0, D)db_i\right\} \times \emptyset(b_i\backslash 0, d)\, p\left(y_{it}\backslash b_i, \beta, \tau, y_{jj}\right) + \sum_{j\neq i} \emptyset(\left(y_{ij}\backslash x_{ij}\beta + z_{ij}b_j, \tau^{-1}\right).\delta_{b_j}$$

After several algebraic operations, we get: **(Kleinman& Ibrahim, 1998)**

$$p\left(b_i\backslash\beta, \tau, y, b^{-i}\right) \propto M|\Sigma_b|^{1/2}|D|^{-1/2}\tau^{ni/2} \times \emptyset(b_i\backslash 0, D)p\left(y_{ij}\backslash b_i, \beta, \tau\right) + \left(\sum_{j\neq i} \tau^{ni/2} exp\left[\frac{-\tau}{2}\left(y_{ij} - x_{ij}\beta - z_{ij}b_j\right)'\left(y_{it} - x_{ij}\beta - z_{ij}b_j\right)\right].\delta_{b\,j}\right)$$

Each term is separated into two elements; the first element is a mixing probability, and the second is a distribution to be mixed, so the second term is the probability of mixing is proportional to: **(Kleinman& Ibrahim, 1998)**

$$\tau^{ni/2}\exp\left(\frac{-\tau}{2}(y_{ij} - x_{ij}\beta - z_{ij}b_j)'(y_{ij} - x_{ij}\beta - z_{ij}b_j)\right)$$

We select from distribution $\delta_{b\,j}$, which means that we set $b_i = b_j$, Also with probability proportional to:

$$M|\Sigma_b|^{1/2}|D|^{-1/2}\tau^{\frac{ni}{2}}\int \exp\left\{\frac{\tau}{2}\left[\left(y_{ij} - x_{ij}\beta\right)'U_i(y_{ij} - x_{ij}\beta)\right]\right\}db_i$$

 P-ISSN 2518-5764
E-ISSN 2227-703X

Where :
$U_i = (\tau z_i \Sigma_b z_i' - I)$
This results in the following distribution, which represents the posterior distribution based on the Dirichlet process As follows:
$$p(b_i \backslash \beta, \tau, y_{ij}) \propto \emptyset(b_i \backslash 0, D) p(y_{ij} \backslash b_i, \beta, \tau, y_{ij}) \qquad (21)$$
And Σ_b the covariance matrix of b which will be found by deriving the posterior distribution in equation (21) as follows:
$$p(b_i | \beta, \tau, D, \{y_{ij}\}) \propto exp\left(\frac{-1}{2} b_i' D^{-1} b_i\right)$$
$$\times\, exp\left(\frac{-1}{2}\sum_{t=1}^{T_i}\left(y_{ij} - x_{ij}'\beta - z_{ij}' b_i\right)^2\right)$$
$$exp\left\{\frac{-\tau}{2}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\left(y_{ij} - x_{ij}'\beta - z_{ij}' b_i\right)^2 + \left(\frac{-1}{2} b_i' D^{-1} b_i\right)\right\}$$
$$exp\left\{\frac{-\tau}{2}\sum_{i=1}^{n}\sum_{j=1}^{T_i}\begin{pmatrix} y_{ij}^2 + \beta' x' x\beta + b' z z' b + 2b' z x'\beta - 2 \\ y x_{ij}'\beta - 2y z_{ij}' b_i\end{pmatrix} + \left(\frac{-1}{2} b_i' D^{-1} b_i\right)\right\}$$
Hence, compare the resulting distribution with the normal distribution to obtain the covariance of the resulting distribution:
$$exp\left\{\frac{-1}{2\sigma^2}\left[\mu^2 + \mu_n^2 - 2\mu\mu_n\right]\right\}$$
$$\frac{b'b}{2\Sigma_b} = \left[\frac{-1}{2} b' D^{-1} b + \frac{-\tau}{2}\sum_{i=1}^{n}\sum_{j=1}^{T} b' z z' b\right]$$
$$\frac{-b'b}{2\Sigma_b} = \frac{-b'b}{2}\left[D^{-1} + \tau\sum_{i=1}^{n}\sum_{j=1}^{T} z z'\right]$$
$$\frac{1}{\Sigma_b} = \left[D^{-1} + \tau\sum_{i=1}^{n}\sum_{j=1}^{T} z z'\right]$$
$$\Sigma_b = \left[D^{-1} + \tau\sum_{i=1}^{n}\sum_{j=1}^{T} z z'\right]^{-1} \qquad (22)$$
Then, obtain the mean:
$$\frac{2\mu b'}{2\Sigma_b} = \frac{-\tau}{2}\left[2b' z x'\beta - 2y b' z\right]$$
$$\frac{\mu b'}{\Sigma_b} = \frac{-\tau}{2} 2b' z\left[x'\beta - y\right]$$
$$\frac{\mu}{\Sigma_b} = -\tau\sum_{i=1}^{n}\sum_{t=1}^{T} z\left[x'\beta - y\right]$$
$$\mu = \Sigma_b\left(-\tau\sum_{i=1}^{n}\sum_{t=1}^{T} z\left[x'\beta - y\right]\right)$$
$$\mu = \Sigma_b\left(\tau\sum_{i=1}^{n}\sum_{t=1}^{T} z\left[y - x'\beta\right]\right)$$

Then the Bayes estimator for the non-parametric is:

$$\widehat{z'_{ij} b_i} = \Sigma_b \left(\tau \sum_{i=1}^{n} \sum_{t=1}^{T} z[y - x'\beta] \right) \quad (23)$$

Covariance Estimation:
If D is assumed to follow the Wishart distribution

$$D^{-1} \sim Wishart\ (v_0, S_0) \quad (24)$$

$$f(D^{-1}) = \frac{|D^{-1}|^{(v_0-p-1)/2}\, e^{-\frac{-1}{2} tr(S_0^{-1} D^{-1})}}{2^{\frac{v_0 p}{2}} |S|^{\frac{v_0}{2}}\, \Gamma p(\frac{v_0}{2})} \quad (25)$$

Then, after getting rid of the constants and multiplying the prior distribution by the likelihood function of b_i, we get: **(west et al., 1994)**

$$p(D^{-1}|\{b_i\})$$
$$\propto |D^{-1}|^{(v_0-p-1)/2} exp\ (\frac{-1}{2} tr(S_0^{-1} D^{-1})$$
$$\times |D^{-1}|^{n/2} exp\left(\frac{-1}{2} \sum_{i=1}^{n} b_i' D^{-1} b_i\right)$$
$$\propto |D^{-1}|^{\frac{v_0+n-1}{2}} exp\left\{\frac{-D^{-1}}{2}\left(S_0^{-1} + \sum_{i=1}^{n} b_i' b_i\right)\right\}$$

Therefore, the posterior distribution will be a Wishart distribution with the following parameters:

$$[D^{-1}|\{b_i\}] \sim Wishart\left(v_0 + n, \left(S_0^{-1} + \sum_{i=1}^{n} b_i b_i'\right)^{-1}\right) \quad (26)$$

2.4.2.2 Bayesian MCMC sampling:

It is one of the techniques of the Monte Carlo Markov Chain and plays an important role in the analysis of the posterior distribution in the Bayesian estimation. The posterior distribution of the model contains all the information related to the prior distribution and the likelihood function and can be used to provide probability data about the parameters.

However, due to the complexity of the studied models, it is difficult to analyze their posterior distributions , This problem can be overcome by applying Markov Chain Monte Carlo (MCMC) techniques, where large samples are drawn from the posterior distributions, and then these samples are used to summarize the posterior distributions. This is done by employing the Gibbs Sampler, where the vector of each parameter is updated by taking its conditional distribution over the rest of the parameters of the other components. After eliminating some of the initial draws, the resulting Markov chains converge for posterior distributions. Sampling continues until the asymptotic posterior distributions are reached. Below is a summary of the Gibbs Sampler algorithm based on posterior distributions :
(**Chen et al., 2000**)(**Robert& Casella, 1999**) (**Tanner &wong, 1987**)

0. Choose starting values for $\tau, \{b_i\}, D^{-1}, \{y_{ij}\}$.
1. Sample β from $[\beta|\{b_i\}, \tau, \{y_{ij}\}]$, which is a Normal distribution.
2. Sample τ from $[\tau|\{b_i\}, \beta, \{y_{ij}\}]$, which is a Gamma distribution.
3. Sample y from $[\{y_{ij}\}|\{b_i\}, \beta, \tau]$, which is a Normal distribution for i=1,….,n, and j=1,….,t.
4. Sample {bi} from $[b_i|\beta, \tau, D, \{y_{ij}\}]$, which is a Normal distribution independently for i=1,….,N.
5. Sample D^{-1} from $[D^{-1}|\ \{b_i\}]$, which is a Wishart distribution.

6. Repeat Steps 1-5 using the updated values of the conditioning variables.

2.4.3 Finding the estimation of the parametric component:

Then the parametric part is estimated after replacing the nonparametric component in model (14) as follows:

$$y_{ij} - x'_{ij}\beta = \widehat{z'_{ij}bi} + \varphi_{ij}$$

$$y^* = x'_{ij}\beta + \varphi_{ij} \tag{27}$$

$$y^* = y_{ij} - \widehat{z'_{ij}bi}$$

Then using the ordinary least squares method as follows:

$$\hat{\beta} = \left(x'_{ij}x_{ij}\right)^{-1}\left(x_{ij}y^*\right) \tag{28}$$

2.4.4 The Simulation :

The R Language program was used to carry out simulation experiments using ten cross sections (n = 10) with three time periods (t = 20, t = 10, t = 5). Thus, we will have three sizes of samples (NT = 50), (NT = 100), and (NT = 200). Each experiment was repeated to obtain accurate and stable results (Replicates = 1000).

2.4.4.1 Generate random variables:

The variables of a semi-parametric random effects model were generated for the panel data as follows:

1- The explanatory variables (S_k) that follow the normal distribution are generated based on the (Box-Muller) method by generating two random variables U_1 and U_2 that follow the uniform distribution U(0,1) and then transforming them into two independent random variables S_1 and S_2 that follow the standard normal distribution According to the following formula:

$$\left.\begin{aligned} S_1 &= (-2Ln\,U_1)^{1/2} * \cos(2\pi U_2) \\ S_2 &= (-2Ln\,U_1)^{1/2} * \sin(2\pi U_2) \end{aligned}\right\} \tag{29}$$

The formula:

$$X = \mu + \sigma * S \tag{30}$$

is employmed to transform the variables from the standard normal distribution to a normal distribution with a mean μ and variance σ^2.

And the process of generating variables takes place according to the number of cross-sections and time periods mentioned in this point. Thus, the semi-parametric model consists of the two variables $X = x_1, x_2$ As well as the explanatory variable z that is generated so that it follows the uniform distribution.

2- The random errors e are generated so that they follow the normal distribution as stated in formula (30), while the vector of random errors for individual effects u_i is generated so that it follows the standard normal distribution as in formula (29) then it is transforming to a normal distribution with a mean of zero and a variance of σ^2 as follows:

$$u_i = \sigma^2 * S \tag{31}$$

Three values of error variance σ^2 have been chosen :

$a)\ \sigma^2 = 1$

$b)\ \sigma^2 = 5$

$c)\ \sigma^2 = 10$

Regarding the error vector v_{ij}, it is directly generated from the generated data and based on the following formula derived from previously published research. (Evdokimov, 2010)

$$v_{ij} = x_{ij}e \quad , e \sim N(0, \sigma^2) \tag{32}$$

Thus, it is possible to obtain the compound error variable φ_{ij} for the random model:

$$\varphi_{ij} = u_i + v_{ij}$$

3- The nonparametric smoothing functions $m(z_{ij})$ that were used in this research are the (linear, quadratic and exponential) functions respectively and in the following formulas : (Wang et al., 2004)

$a)\ m_1(z_{ij}) = z$ (33)
$b)\ m_2(z_{ij}) = 3.2\ z^2 - 1$ (34)
$c)\ m_3(z_{ij}) = \exp\{-32(z-0.5)^2)\}$ (35)

4- As for the dependent variable y_{ij}, it is generated directly using explanatory variables, random errors, and smoothing functions that were generated in paragraphs 1 to 3 above, according to the needed model.
5- The parameter values of the parametric component of the semi-parametric model were also determined by estimating them using the least squares method in a way that is consistent with the nature of the standard data studied in the applied side, as follows:
$\beta_0 = 0.2\ , \beta_1 = -0.09\ , \beta_2 = 0.1$
6- The Gaussian kernel function was used, in addition to the use of the cross-validation method for estimating the bandwidth h, for all kernel estimation methods, and the generalized cross-validation method for estimating the smoothing parameter λ, for all splines estimation methods, according to the steps that were presented in the theoretical side.

Table (1) describes the models estimated according to the estimation methods

Table1: describes the methods estimated

No.	**Model**	**The method estimation**
1	Method I	Speckman Method
2	Method II	Bayes Method

3. discussion of Results :

The values of the Average Mean Squared Error(AMSE) for different models for different levels of variance and for different sample sizes using the nonparametric smoothing functions in equations (29,30 and 31) are presented in the tables from (2-4) as follow:

Table2: Values of Average Mean Squared Error (AMSE) by using Linear function in equation(33)

σ_e^2	**Method**	**AMSE**		
		n=50	**n=100**	**n=200**
$\sigma_e^2 = 1$	Method I	0.19082856	0.209704699	0.214879061
	Method II	0.162442607	0.179088110	0.199329073
$\sigma_e^2 = 5$	Method I	2.4759138	2.98611972	3.0214238
	Method II	1.3284889	1.50316086	1.7163896
$\sigma_e^2 = 10$	Method I	9.3894085	10.3602578	10.97792714
	Method II	18.7009065	18.7559739	28.845967

Table3: Values of Average Mean Squared Error (AMSE) by using Quadratic function in equation(34)

σ_e^2	**Method**	**AMSE**		
		n=50	**n=100**	**n=200**
$\sigma_e^2 = 1$	Method I	0.172637064	0.185912097	0.199814385
	Method II	0.152887581	0.17447287	0.196816504
$\sigma_e^2 = 5$	Method I	2.3737339	2.77102026	2.8805216
	Method II	1.3951073	1.4082165	1.452309144
$\sigma_e^2 = 10$	Method I	9.38487323	9.8833582	10.8376703

	Method II	10.548904	16.3183375	23.6989039

Table4: Values of Average Mean Squared Error (AMSE) by using Exponential function in equation(35)

σ_e^2	**Method**	**AMSE**		
		n=50	**n=100**	**n=200**
$\sigma_e^2 = 1$	Method I	0.1169552668	0.123762276	0.160942112
	Method II	0.11476246	0.120329479	0.157833968
$\sigma_e^2 = 5$	Method I	2.0337531	2.11391461	2.385008261
	Method II	0.99439412	1.17621229	1.34821452
$\sigma_e^2 = 10$	Method I	8.0384536	9.5718134	10.25457564
	Method II	10.3424103	10.6181225	13.19540385

The values of Average Mean Squares Error in Tables (2,3,4) showed the following results:

1- The models estimated according to the Bayes method (Model II) gave average mean square error values at the variance levels (1,5) less than those provided by the Speckman method (Model I).

2- At variance levels (10) the models estimated using the Speckman method (Model I) yielded average mean square error values that were lower than those obtained using the Bayes method (Model II).

3- The values of the average mean square error estimated increase when increasing the level of variance for the two methods.

4- The average mean square error values increase with increasing sample size.

4. Conclusions:

After the simulation experiments were carried out and the results presented and analyzed, the researcher concluded the following :

1- When increasing the variance level increases the average mean square error values. Therefore, the relationship between the variance level and the efficiency of the estimation methods is inverse.

2- The Bayes method be more efficient efficient at low variance levels, While the Speckman method be more efficient at high variance levels.

3- Increasing the sample size leads to an increase in the mean square error of the two methods, which means that the relationship between sample size and the efficiency of the mean square error is inverse.

4- Using the high variances makes it easy for us to determine which method is more efficient compared to the other.

5- The average mean square error values for the two methods employing exponential smoothing functions were less than those employing quadratic and linear smoothing functions.

Authors Declaration:

Conflicts of Interest: None

-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, Which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.

- Ethical Clearance: The Research Was Approved By The Local Ethical Committee in The University.

References:

1.Abd-Alhafez,A. and Rashid, D. (2013), Comparison Robust M Estimate With Cubic Smoothing Spline For time –Varying Coefficient Model For Balance Longitudinal Data. Journal of Economics and Administrative Sciences, Vol 19,No.72,pp:398-413.

2.Abd-Alrazzaq, Ali(2015), Estimating Missing Data In Panel data Model with Practical Application. Master's thesis, University of Baghdad, Baghdad.

3.Abd-Alreda, M.A. and Fadam,E. (2019), Nonparametric and Semiparametric Bayesian Estimation in survival function analysis, Journal of Economics and Administrative Sciences, Vol 25,No.112,pp:461-480.

4.Burhan,Y. and Hammoud,M. (2018), Comparison between the methods estimate nonparametric and semiparametric transfer function model in time series the Using simulation. Journal of Economics and Administrative Sciences, Vol 24,No.106,pp:375-391.

5.Castelein, A., Fok, D., and Paap, R. (2020), Heterogeneous Variable Selection in Nonlinear Panel data Models: A Semi-Parametric Bayesian Approach, (No. TI 2020-061/III). Tinbergen Institute Discussion Paper.

6.Conklin, M. (2001), Monte Carlo Methods in Bayesian Computation, Technometrics, Vol 43,No.2, pp:240-240.

7.Ferguson, T. S. (1973), A Bayesian analysis of some nonparametric problems. The annals of statistics, pp:209-230.

8.Green. W. (2002), Econometric Analysis. 5ed, Prentice Hall, New Jersey.

9.Hamza , R. K. and Jaafar, M. (2018), Comparison of estimations methods of the entropy function to the random coefficients for two models: the general regression and Swamy of the panel data", Journal of Economics and Administrative Sciences, Vol 25,No.110,pp:371-391.

10. Hardle,W., Muller, M., Sperlich, S., and Werwatz, A. (2004), Nonparametric and semiparametric models. Berlin: Springer, Vol. 1.

11. Jochmann, M., and León-González, R. (2004), Estimating the demand for health care with Longitudinal data: a semiparametric Bayesian approach. Health Economics, Vol 13,No.10, pp:1003-1014.

12. Kamel, R. and Aboudi,E. (2021), Comparison Of Some Methods For Estimating a Semi-parametric Model For Longitudinal Data, Journal of Economics and Administrative Al-Mustansiriya, Vol 28,No.127,pp:261-249.

13. Khalil, Y. and Fadam,E. (2016), Compare to the conditional logistic regression models with fixed and mixed effects for longitudinal data, Journal of Economics and Administrative Sciences, Vol 23,No.98,pp:406-429.

14. Kleinman, K. P., and Ibrahim, J. G. (1998), A semiparametric Bayesian approach to the random effects model. Biometrics, PP: 921-938.

15. Laird, N. M. and Ware, J. H. (1982), Random-effects models for longitudinal data. Biometrics No.38, pp:963-974.

16. Li, Q., and Stengos, T. (1996), Semiparametric estimation of partially linear Longitudinal data models. Journal of econometrics, Vol 71,No.1, pp:389-397.

17. Liu, S., You, J., and Lian, H. (2017), Estimation and model identification of longitudinal data time –varying nonparametric models. Journal of Multivariate Analysis, No.156,pp:116-136.

18. Mohaisen, A. J., & Abdulhussein, A. M. (2014). Fuzzy sets in semiparametric Bayes Regression. Basrah Journal of Science (A), 32(1), 141-167.

19. Nauef Al-Qazaz, Q. N., and Ali, A. H. (2022), Bayesian inference of fractional brownian motion of multivariate stochastic differential equations. International Journal of Nonlinear Analysis and Applications, Vol 13,No.1, pp:2425-2454.

20. Nayef Al-Qazaz, Q. N., and Shawkat, L. N. (2022), Using some methods to estimate the parameters of the Multivariate Skew Normal (MSN) distribution function with missing data. International Journal of Nonlinear Analysis and Applications, Vol13,No.1, pp:2333-2350.

21. Robert, C. P., Casella, G., and Casella, G. (1999), Monte Carlo statistical method. New York: Springer. Vol. 2 .

22. Robinson, P. M. (1988). Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society, No.56, pp:931-954.

P-ISSN 2518-5764
E-ISSN 2227-703X

23. Rodriguez-Poo,J.M., and Sobern,A.(2017), Nonparametric and semiparametric Longitudinal data models:Recent developments. Journal of Economic Surveys, vol31,No.4, pp:923-960.
24. Sadiq, N. J.(2015), Estimate the Regression Model of Longitudinal Data with Drop-outs in Response Variable with Application in Medical field.Ph.D. thesis, University of Baghdad, Baghdad.
25. Shaker,A. and Nayef,Q. (2016), Comparison Semiparametric Bayesian Method with Classical Method for Estimating Systems Reliability using Simulation Procedure. Journal of Economics and Administrative Sciences, Vol 23,No.97,pp:378-393.
26. Speckman, P. (1988). Kernel smoothing in partial linear models. Journal of the Royal Statistical Society: Series B (Methodological), Vol.50, No.3, pp:413-436.
27. Tanner MA, Wong W. ,(1987), The calculation of posterior distributions by data augmentation (with discussion). Journal Amer Statist Assoc ; No.82: pp:528–550.
28. Tukey, J. W. (1961), Curves as parameters, and touch estimation. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. Volume 1: Contributions to the Theory of Statistics, University of California Press, No. 4, pp: 681-695.
29. Wang, H. (2014), Essays on Semiparametric Ridge-Type Shrinkage Estimation, Model Averaging and Nonparametric Panel Data Model Estimation. University of California, Riverside.
30. West, M., Muller, P., and Escobar, M. D. (1994), Hierarchical priors and mixture models, with applications in regression and density estimation. In Aspects of Uncertainty: A Tribute to D. V. Lindley, A. F. M. Smith and P. R. Freeman (eds). London: Wiley.
31. Zeger, S. L., Diggle, P. J., Heagerty, P. and Liang, K. Y., (2002), Analysis of longitudinal data. Second edition, oxford university press, New York, No.69.

Journal of Economics and Administrative Sciences
2024; 30(142), pp. 449-464

P-ISSN 2518-5764
E-ISSN 2227-703X

مقارنة بين طريقة تقدير سبيكمان وبيز لنموذج البيانات الطولية المتوازنة شبه المعلمي

احمد عبد الصمد حبيب[1]
جامعة بغداد / كلية الادارة والاقتصاد / قسم الاحصاء
العراق
Ahmed.Thamer1101@ coadec.uobaghdad.edu.iq

قتيبة نبيل نايف[2]
جامعة بغداد / كلية الادارة والاقتصاد / قسم الاحصاء
العراق
dr.qutaiba@coadec.uobaghdad.edu.iq

Received:16/9/2023 Accepted: 6/11/2023 Published Online First: 30 /8/ 2024

مستخلص البحث:

تهدف هذه الدراسة إلى استخدام الانحدار شبه المعلمي لأنموذج البيانات الطولية المتوازنة، حيث تعاني نماذج الانحدار المعلمية من مشكلة القيود الصارمة، بينما تعاني نماذج الانحدار اللامعلمية على الرغم من مرونتها من مشكلة تعدد الابعاد. وبالتالي يعد الانحدار شبه المعلمي حلاً مثاليًا للتخلص من المشكلات التي يعاني منها الانحدار المعلمي واللامعلمي. والميزة الكبيرة لهذا النموذج هو أنه يحتوي على كافة السمات الإيجابية التي تضمنها النموذجان السابقان، مثل احتوائه على قيود صارمة في مكونه المعلمي، والمرونة الكاملة في مكونه اللامعلمي، ووضوح التفاعل بين مكوناته المعلمية واللامعلمية.

وبناء على ما سبق، تم استخدام طريقتين لتقدير نموذج البيانات الطولية المتوازنة شبه المعلمية. الأول هو طريقة التقدير البيزية؛ والثاني هو طريقة سبيكمان، التي تقدر دالة التمهيد اللامعلمية المجهولة من خلال استخدام طريقة تمهيد النواة (kernel) نادرايا-واتسون. وكان الهدف هو إجراء مقارنة بين طريقة التقدير بيزي وطريقة التقدير التقليدية المتمثلة بطريقة سبيكمان. تم استخدام ثلاثة أحجام مختلفة للعينات في تجارب المحاكاة: 50، 100، 200. أظهرت نتائج الدراسة أن طريقة التقدير البيزي هي الأفضل عند مستويات التباين المنخفضة (1،5)، في حين أن طريقة سبيكمان هي الأفضل عند مستويات التباين العالية (10).

نوع البحث: بحث مستل من اطروحة دكتوراه
المصطلحات الرئيسة للبحث: بيز، سبيكمان، نادرايا- واتسون، شبه المعلمي، البيانات الطولية المتزنة.