Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 5, No. 1, 2023

Research on Two-stage Estimation of Partially Linear Single-index Model with Longitudinal Data

Chaojie Chang

School of Mathematics and Science, China University of Geosciences (Wuhan), Wuhan, China

Abstract: Partial linear single-index model is a kind of semi-parametric model with wide application. In this paper, we deal with the partial linear single-index model under longitudinal data. A "two-stage estimation method" without iteration by using local polynomial and bias correction generalized estimation equation is proposed. under some regularity conditions, the asymptotic properties of the connection function and unknown parameter estimator are investigated. Numerical simulation shows that the proposed method is robust.

Keywords: Partial linear single-index model, Longitudinal data, Local polynomial, Generalized estimation equation for correcting deviation, Asymptotic normality.

1. Introduction

This paper studies the partial linear single-indicator model under the following longitudinal data:

$$Y_{ij} = Z_{ij}^T\theta + \mathrm{g}\left(X_{ij}^T\beta\right) + \varepsilon_{ij}, i = 1,\cdots,n, j = 1,\cdots,m_i \quad (1)$$

Among them, the explanatory variables X and Z are related to the model error ε Independent from each other. g(·)is a known nonparametric connection function, β is an unknown q-dimension index parameter, θ is an unknown p-dimensional linear parameter, ε_{ij} is the mean value is 0, and the variance is $0 < \sigma^2 < \infty$ random error. To ensure the identifiability of the single indicator part, suppose $\|\beta\| = 1$, and the first non-zero element of the parameter vector is positive. In this paper, under the longitudinal data, each individual is observed for a limited number of times, namely $m_i < c$, then the total number of observations is $= \sum_{i=1}^{n} m_i$ Where m_i is a bounded positive integer sequence, and it can be assumed that n is infinite, which means that the total number of observations N and the total number of subjects n are of the same order, that is, $N \to \infty$ and $n \to \infty$ are equivalent.

Partial linear single-indicator model is the combination of linear model and single-indicator model. Its application range is very wide, which has aroused the research interest of many scholars. Many methods have been proposed to estimate unknown parameters and non-parametric connection functions[1-4]. When $q = 1$, the model (1) is a partial linear model of longitudinal data [5-6]; When $p = 0$, the model (1) is a single indicator model of longitudinal data [7-8]; When g(·) selects 0, model (1) is a linear regression model of longitudinal data [9].

The existing estimates of partial linear single-indicator models are mainly based on the mean regression of least squares or likelihood method, and the distribution of random error needs to be assumed. However, if the assumption is incorrect, the estimation results may be inaccurate or even wrong, and the research work will be meaningless. In addition, at present, the linear parameters in the model θ_0 and index parameters β_0 is almost estimated at the same time, and requires multiple iterations. The calculation is very complex and time-consuming. Then, due to β_0 and θ_0 may have some correlation, which may cause β_0 is difficult to identify, which will also lead to inaccurate estimation results.

In view of the above problems, without considering the intra-group correlation, this paper proposes a "two-stage estimation method" without iteration based on local polynomial estimation and bias correction generalized estimation equation β And linear parameters θ This method can reduce the running time and improve the operation efficiency. In addition, in order to solve β_0 and θ_0 For the linear correlation between X and Z, the formula (2) is introduced to eliminate the correlation between X and Z, so as to improve the accuracy of parameter estimation.

In section 1, we give a "two-stage estimation method" without iteration based on local polynomial estimation and bias correction generalized estimation equation [10]. In section 2, we study the index parameters under some regularity assumptions β And linear parameters θ The asymptotic normality of the estimator and the asymptotic normality of the estimator of the connection function g(·). The random simulation experiment in section 3 shows that the estimator obtained by this method is robust.

Assume that the observation value is $\{(X_{ij},Y_{ij},Z_{ij}); i = 1,\cdots,n, j = 1,\cdots,m_i\}$ from the sample of model (1). For the convenience of calculation：

$$Y_i = \left(y_{i1},\cdots,y_{im_i}\right)^T, X_i = \left(x_{i1},\cdots,x_{im_i}\right)^T$$
$$Z_i = \left(z_{i1},\cdots,z_{im_i}\right)^T, \varepsilon_i = \left(\varepsilon_{i1},\cdots,\varepsilon_{im_i}\right)^T$$

Then the model (1) can be written as the following vector matrix form:

$$Y_i = Z_i^T\theta + g\left(X_i^T\beta\right) + \varepsilon_i, i = 1,\cdots,n$$

To ensure the identifiability of the model, it is assumed that $\beta = \left(1,\beta_1,\cdots,\beta_{p-1}\right)^T$. In order not to lose generality, real parameters are assumed β_0 is a positive definite matrix. In addition, based on the linear correlation between X and Z,let

$$Z_i = \varphi\left(X_i^T\beta_z\right) + \eta_i \quad (2)$$

here $\varphi(\cdot)$ is an unknown function from R^d to R^q, β_z is a $p \times d$ with standard orthogonal columns matrix, η the mean value of is zero and is independent of X. The dimension d is usually much smaller than the dimension p of X,

which is a common dimensionality reduction assumption in the literature. In this chapter, we conduct statistical inference research under the condition of $d = 1$. Generally, once you get β_0The $\sqrt{n}$ of is estimated and inserted (1), and the best estimate can be achieved by the method developed for the partial linear model θ_0, however, β_0 and θ_0 may be related, causing β_0 is difficult to identify. This is the advantage of introducing model assumption (2), because it allows deleting the Z part related to X, so that the residual in (2) η Will be independent of X. Similarly, it is necessary to add an identifiability condition, i.e. $\|\beta_z\| = 1$ and the first component of the parameter is positive.

The parameter vector is given below β_0 and θ_0 and connection function g(·) The estimation algorithm of "two-stage estimation method":

Algorithm for Stage One:

1. First, construct a regression model Z_{ij} and X_{ij} is regressed to obtain β_z estimator of $\hat{\beta}_z$

2. Using local smoothing estimation to get $\hat{\eta}_{ij} = Z_{ij} - \varphi(X_{ij}^T\hat{\beta}_z)$,so $Y_{ij} = \hat{\eta}_{ij}\theta_0 + h(X_{ij}^T\beta_0 + X_{ij}^T\hat{\beta}_z) + \varepsilon_{ij}$, her η and X are independent of each other.

3. For $Y_{ij}$and $\hat{\eta}_{ij}$ carries out linear regression to obtain θ_0 initial estimate of $\hat{\theta}_0$

4. Construct a regression model $Y_{ij} - Z_{ij}^T\hat{\theta}_0$与 and X_{ij} is regressed to obtain β_0 initial estimate of $\hat{\beta}_0$

5. Using local polynomial smoothing estimation, the initial feasible estimator of the connection function g(·) and its first derivative g′(·) is obtained ĝ and ĝ′

$$\hat{g}(u) = \hat{g}(u;\theta,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} W_{nij}(u;\hat{\beta}_0)(Y_{ij} - Z_{ij}^T\hat{\theta}_0)$$

$$\hat{g}'(u) = \hat{g}'(u;\theta,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \widetilde{W}_{nij}(u;\hat{\beta}_0)(Y_{ij} - Z_{ij}^T\hat{\theta}_0)$$

Algorithm for Stage Two:

6. Use Step 4 to get the initial estimate $\hat{\beta}_0$, obtained by solving the deviation correction generalized estimation equation θ_0 initial estimate of $\hat{\theta}$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij} - Z_{ij}^T\theta - \hat{g}(X_{ij}^T\beta)\}\{Z_{ij} - \hat{E}(Z_{ij}|X_{ij}^T\beta)\} = 0$$

7. Use the updated θ_0 initial estimate of $\hat{\theta}$ form a new residual $Y_{ij} - Z_{ij}^T\hat{\theta}$,Then it is obtained by solving the deviation correction generalized estimation equation as follows β_0 initial estimate of $\hat{\beta}$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{Y_{ij} - Z_{ij}^T\theta - \hat{g}(X_{ij}^T\beta)\}g'(X_{ij}^T\beta)\{X_{ij} - \hat{E}(X_{ij}|X_{ij}^T\beta)\} = 0$$

8. Use the updated estimates of $\hat{\theta}$ and $\hat{\beta}$ in steps 6 and 7 to updated the estimate of ĝ and ĝ′,following the steps as described in step 5.

The first stage is to use the linear regression method and local linear smoothing estimation to obtain the initial estimates of unknown parameters, connection functions and their derivatives and residuals. In the second stage, the final estimates of unknown parameters and connection functions are obtained by using the deviation correction generalized estimation equation. The algorithm can obtain the asymptotic property of the unknown parameter estimator without iteration, and solve β_0 and $\theta_0$may be a related problem that leads to inaccurate parameter estimation; In addition, the estimators of the connection function g(·) and its derivative ĝ′(·) are also obtained.

2. Main Result

In order to study the theoretical results of the estimators obtained by the "two-stage estimation method", the following regularity assumptions are first given:

C1 Individuals and individuals are independent and equally distributed.

C2 (i) The distribution of X_{ij} has a compact support set $\mathcal{A}$;

(ii) The bounded positive function $f(t)$is $X_{ij}^T\beta$ density function on T, and β_0 The domain of satisfies the Lipschitz condition of order 1.here $T = \{t = X_{ij}^T\beta; X_{ij} \in A, i = 1,\cdots,n, j = 1,\cdots,m_i\}$.

C3 (i) Join functions g and g_{1i} has a second order continuous partial derivative, where$g_{1i}$is the function vector g_1 the ith component of i, $1 \le i \le q$,$g_1(t) = E(Z_{ij}|X_{ij}^T\beta = t)$;

(ii) g_{2j}satisfies the first-order Lipschitz condition, where g_{2j} is the jth component of g_2, $1 \le j \le p$,$g_2(t) = E(X_{ij}|X_{ij}^T\beta = t)$.

C4 On R^1, the kernel function $K(\cdot)$ is a continuous bounded probability density function that satisfies the Lipschitz condition and satisfies

$$\int_{-\infty}^{\infty} u^2K(u)du \neq 0, \int_{-\infty}^{\infty} |u|^2K(u)du < \infty.$$

C5 (i) it has a constant $M > 0$, so that for anyi,j satisfies $\sup_{t\in T} E\left(\|Z_{ij}\|^2 \middle| X_{11}^T\beta = t\right) \le M < \infty$;

(ii) Random error ε_{ij} independent of covariate X_{ij} and Z_{ij}, and there is a constant $M > 0$, $(\varepsilon) = 0$,$0 < var(\varepsilon) = \sigma^2 < \infty$,$E(\varepsilon^4) \le M < \infty$.

C6 When $n \to \infty$,bandwidth sequence h and h_1 satisfie

(i) $nh^2/\log^2 n \to \infty$, $\lim_{n\to\infty} nh^5 < \infty$;

(ii) $nh^2h_1^3/\log^2 n \to \infty$,$nh^4 \to \infty$.

C7 Definition $Y^* = Y_{ij}^* - Z_{ij}^T\theta_0$, Σ is a bounded positive definite matrix

$$\Sigma = Cov(Z - E(Z|X^T\beta_0))$$

$$B_0 = E\{g'(X^T\beta_0)^2[X - (X|X^T\beta_0)][X - (X|X^T\beta_0)]^T\}$$

$$B_1 = E\{f_{Y^*}g(X^T\beta_0)g'(X^T\beta_0)^2[X - (X|X^T\beta_0)][X - (X|X^T\beta_0)]^T\}$$

Condition (C1) is the assumption of independence. See reference [11] for details. Lipschitz condition and standard smoothness condition in condition (C2) and condition (3). For some common regularity and slip assumptions of conditional (C4) kernel functions, to ensure that the theoretical results are well established. The condition (C5) is to satisfy the existence of the second moment, so that the proposed parameter and the single exponential function estimator are consistent and asymptotically normal. Condition (C6) is the bandwidth h used to estimate the connection function g(·), and the other band width h_1 is to control the change of ĝ′(·), see reference [12]. The condition (C7) ensures that the variance limit of the parameter estimator exists.

Let $\rho_l = \int_{-\infty}^{\infty} u^lK(u)du$,$\varrho_l = \int_{-\infty}^{\infty} K^l(u)du$,$l = 1,2,3$, the following are the connection function g(·) and unknown parameter components θ and β Results of asymptotic

properties of.

Theorem 1 Assume that the above regularity and smoothness conditions C1-C4 hold, and if $nh^5 \to 0$, the initial estimator$\hat{\beta}_0,\hat{\theta}_0$ satisfied $\|\hat{\beta}_0 - \beta_0\| = O_p(n^{-1/2}),\|\hat{\theta}_0 - \theta_0\| = O_p(n^{-1/2})$, then when $n \to \infty$

$$\sqrt{nh}\{\hat{g}(u;\hat{\beta},\hat{\theta}) - g(u) - h^2 b(u)\} \xrightarrow{D} N(0,A(u))$$

here $b(u) = \frac{1}{2}g''(u)\rho_2$, $A(u) = \frac{\varrho_2}{f(u)[f_{Y^*}(g(u)|u)]^2}$

Theorem 2 Assume that the above regularity and smoothness conditions C1-C7 hold, and the initial estimator $\hat{\beta}_z,\hat{\beta}_0$ satisfied $\|\hat{\beta}_z - \beta_z\| = O_p(n^{-1/2}),\|\hat{\beta}_0 - \beta_0\| = O_p(n^{-1/2})$, then when $n \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0,\sigma^2\Sigma^{-1})$$

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0,\sigma^2 B_1^{-1}B_0B_1^{-1})$$

where B_1^{-1} is the inverse of B_1 ,the definition of Σ、B_0 and B_1 see C7.

3. Simulation Study

In this section, this paper considers the effectiveness of the "two-stage estimation method" by applying the Monte Carlo method in the case of limited samples.

First, generate some random numbers in the following model:

$$Y_{ij} = Z_{ij}^T\theta + g(X_{ij}^T\beta) + \varepsilon_{ij}, i = 1,\cdots,n, j = 1,\cdots,m_i$$

where Z_{ij} is a covariate, a $0-1$ distribution with a parameter of 0.5, $X_{ij} = (X_{i1},X_{i2},X_{i3},X_{i4},X_{i5})^T$ X_{i1}、X_{i2}、X_{i3}、$X_{i4}$and X_{i5} are independent of each other and are uniform distribution from the $(0,1)$ interval. During the simulation experiment, errors $\varepsilon_{ij} = (\varepsilon_{i1},\varepsilon_{i2},\varepsilon_{i3},\varepsilon_{i4},\varepsilon_{i5})^T, i = 1,2,3,4,5$ obey standard normal distribution. $\beta_z = (0.5,0,0.5,0.5,-0.5)^T, \beta_0 = (0.75,0.5,-0.25,-0.25,0.2)^T$ and $\theta_0 = 1$ the connection function $g(u) = \sin\left(\frac{\pi(u-C)}{A-C}\right)$, $A = \frac{\sqrt{3}}{2} - \frac{1.645}{12}$, $C = \frac{\sqrt{3}}{2} + \frac{1.645}{12}$.

In addition, this paper uses the kernel function $K_h(x) = \frac{2}{h\sqrt{2\pi}}exp\left(-\frac{(x-\mu)^2}{2h^2}\right)$, h is the window width, which is selected by the following generalized cross validation method and meets the assumption C6:

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\frac{\left(Y_{ij} - Z_{ij}^T\hat{\theta} - \hat{g}_h(X_{ij}^T\hat{\beta})\right)}{\{n^{-1}tr(I - S_h)\}^2}$$

For the convenience of comparison, n is selected as 50 "," 150 and 200 respectively, and the simulation results of each case are based on 500 repeated experiments. In this paper, the following evaluation indicators are used to evaluate the accuracy of parameter estimation: linear parameter θ Estimator θˆ Deviation, standard deviation, mean square error and index parameters of β Estimator βˆ The deviation, standard deviation and mean square error of, and the estimator of the linking function g (·) gˆ (·) The mean and standard deviation of RMSE, where the estimated RMSE of the connection function g (·) is calculated by the square root of the following mean square error:

$$\text{RMSE}(\hat{g}) = \left(\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left[\hat{g}(X_{ij}^T\hat{\beta}) - g(X_{ij}^T\beta)\right]^2\right)^{\frac{1}{2}}$$

Table 1 below shows the deviation, standard deviation and mean square error of the estimator when n is taken as 50 "," 150 "and 200 respectively. Figure 1 shows the scatter diagram of the connection function when n=50", "150" and 200 ".

Table 1. β_0 and θ_0 deviation, standard deviation and mean square error of estimator, g ˆ(·) Mean and standard deviation of RMSE

n	150				200				300			
	bias	**std**	**ste**	**mse**	**bias**	**std**	**ste**	**mse**	**bias**	**std**	**ste**	**mse**
θ	0.0035	0.1274	0.1278	0.0163	0.0024	0.1026	0.1131	0.0125	0.0017	0.0894	0.0921	0.008
β_1	0.0432	0.2644	0.251	0.0718	0.0466	0.2372	0.2163	0.0584	0.0311	0.1918	0.1746	0.0377
β_2	0.0209	0.1304	0.1217	0.0174	0.0193	0.1102	0.1053	0.0125	0.0125	0.0898	0.0854	0.0083
β_3	0.0414	0.2144	0.1996	0.0477	0.0413	0.1956	0.1764	0.0506	0.0264	0.1611	0.1443	0.0267
β_4	0.0221	0.2479	0.2253	0.0143	0.0109	0.2196	0.1945	0.0105	0.0311	0.1761	0.0763	0.0068
β_5	0.0418	0.2005	0.1793	0.0419	0.0432	0.1002	0.0939	0.0352	0.0138	0.0815	0.1302	0.0225
	me		**se**		**me**		**se**		**me**		**se**	
g	0.0412		0.0142		0.0333		0.0121		0.0296		0.0076	

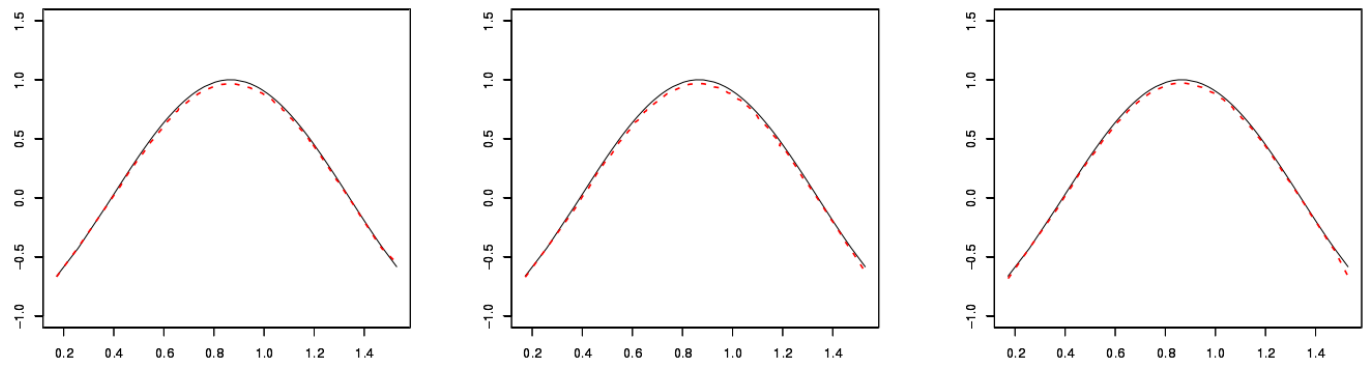

Figure 1. Scatter diagram of real connection function and estimated connection function when n=50,100,200

From Table 1 and Figure 1, the following conclusions can be drawn:

The deviation of parameter estimation, standard error, mean and mean square error of standard error, as well as the mean and standard error of RASE estimated by the connection function all decreased significantly with the

increase of samples. Therefore, the "two-stage estimation method" proposed in this paper is relatively stable in the estimation of parameters, and the connection function estimation and the fitting effect of the real curve are good.

4. Conclusion

This paper presents a new "two-stage estimation method", which is based on local polynomial and bias correction generalized estimation equation. It can estimate the index parameters and linear parameters in turn, and can obtain the asymptotic normality of the estimator. The Monte Carlo simulation results show that the algorithm has good robustness.

References

[1] Shakhawat Hossain and Le An Lac. Optimal shrinkage estimations in partially linear single-index models for binary longitudinal data[J]. TEST, 2021, 30(4) : 1-25.

[2] Quan Cai and Suojin Wang. Inferences with generalized partially linear single-index models for longitaudinal dta[J]. Journal of Statistical Planning and Inference, 2018, 200 : 146-160.

[3] Gaorong Li and Peng Lai and Heng Lian. Variable selection and estimation for partially linear single-index models with longitudinal data[J]. Statistics and Computing, 2015,25(3) : 579-593.

[4] Peng Lai and Gaorong Li and Heng Lian. Quadratic inference functions for partially linear single-index models with longitudinal data[J]. Journal of Multivariate Analysis, 2013, 118 : 115-127.

[5] Zeger S L, Diggle P J. Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. Biometrics, 1994, 50: 689–699.

[6] Yu Ying Jiang. Empirical Likelihood Inference for a Partially Linear Model under Longitudinal Data[J]. Applied Mechanics and Materials, 2013, 2545(353-356) : 3355-3358.

[7] Hongmei Lin et al. A new local estimation method for single index models for longitudinal data[J]. Journal of Nonparametric Statistics, 2016, 28(3) : 644-658.

[8] Peng Lai and Gaorong Li and Heng Lian. Semiparametric estimation of fixed effects panel data single-index model[J]. Statistics and Probability Letters, 2013, 83(6) : 1595-1602.

[9] Ruiqin Tian and Liugen Xue. Generalized empirical likelihood inference in partial linear regression model for longitudinal data[J]. Statistics, 2017, 51(5) : 988-1005.

[10] Xue L G, Zhu L. Empirical likelihood for single-index models. J Multivariate Anal, 2006, 97: 1295–1312.

[11] Chen J, Li D, Liang H, et al. Semiparametric GEE analysis in partially linear single-index models for longitudinal data. Ann Statist, 2015, 43: 1682–1715.

[12] Pang Z, Xue L. Estimation for the single-index models with random effects. Comput Statist Data Anal, 2012, 56: 1837–1853.