Open Journal of Applied Sciences, **2013, 3, 44-48**
doi:10.4236/ojapps.2013.31B1009 Published Online April 2013 (http://www.scirp.org/journal/ojapps)

Imputed Empirical Likelihood for Varying Coefficient Models with Missing Covariates

Peixin Zhao
Department of Mathematics, Hechi University, Guangxi Yizhou, China
Email: zpx81@163.com

Received 2013

ABSTRACT

The empirical likelihood-based inference for varying coefficient models with missing covariates is investigated. An imputed empirical likelihood ratio function for the coefficient functions is proposed, and it is shown that iis limiting distribution is standard chi-squared. Then the corresponding confidence intervals for the regression coefficients are constructed. Some simulations show that the proposed procedure can attenuate the effect of the missing data, and performs well for the finite sample.

Keywords: Empirical Likelihood; Varying Coefficient Model; Missing Covariate

1. Introduction

In practice, missing data frequently occur in many application literatures, and the literatures on statistical analysis of data with missing values have been flourished in the past decade. Parametric regression models with missing data have been widely discussed (see [1,2]). In many practical situations, however, the parametric regression models are not flexible enough to capture the underlying relation between the response and the associate covariates. Hence, Wang [3] and Liang et al. [4] considered the statistical inferences for the partially linear model with missing covariates, which is a useful extension of the parametric regression model. In addition, the following varying coefficient model is another useful extension of the parametric regression model, which has more implements and stronger explanations than the parametric regression model. This paper aims to present an imputed empirical likelihood method for analyzing the varying coefficient model with covariate data missing at random.

Consider the following varying coefficient model

$$Y = X^T\theta(U)+\varepsilon \tag{1}$$

where Y is the response variable, X is the $p\times 1$ covariate vector, U is the scalar covariate, and $\theta(u)=\left(\theta_1(u),\cdots,\theta_p(u)\right)^T$ is a vector of unknown smooth functions. The error ε has mean zero conditional on X and U. In this paper, we focus mainly on the case that the covariate X may be missing at random. That is, the available incomplete data with the sample size of n are denoted as

$$\left(\delta_i, X_i, Y_i, U_i\right),\ i=1,2,\cdots,n$$

where $\delta_i=0$ if X_i is missing, otherwise $\delta_i=1$, and it satisfies that

$$P\left(\delta_i=1\middle|X_i,Y_i,U_i\right)=P\left(\delta_i=1\middle|Y_i,U_i\right)\equiv\pi\left(Z_i\right), \tag{2}$$

where $Z_i=\left(Y_i,U_i\right)$. The supposition (2) is commonly used in the literature of missing data (see [2-5]). It is well known that, in the presence of missing data, the complete case analysis often generate a considerable bias and lose efficiency. Then, it is important to develop some new methods which can take the partially incomplete data into account.

In this paper, an imputed empirical likelihood procedure is proposed to study model (1) under missing covariates. The proposed method can use the information of the incomplete data efficiently, and the limiting distribution of the proposed empirical log-likelihood ratio function is shown to be standard chi-squared. Then the corresponding confidence intervals of the regression coefficients are constructed. Some simulations show that the proposed procedure can attenuate the effect of missing data, and performs well for finite sample.

Compared with the Wald-type confidence intervals, the empirical likelihood based confidence intervals possess several attractive features such as the circumvention of asymptotic variance estimation and the flexible shapes of the confidence intervals determined by data (see [6]). This paper provides an additional positive result of the empirical likelihood inferences varying coefficient models with missing data, which extends the application literature of the empirical likelihood method.

[*] This research was supported by the National Natural Science Foundation of China (Grant No. 11101119).

2. Methodology and Main Results

Let

$$\begin{aligned}\Phi_u(z) &= E\left(X_i\left(Y_i - X_i^T\theta(u)\right)\middle|Z_i = z\right) \\ &= Y_i g_1(z) - g_2(z)\theta(u)\end{aligned},$$

where $g_1(z) = E(X|Z = z)$ and $g_2(z) = E\left(XX^T|Z = z\right)$. Then, by a simple calculation, we have that

$$E\left\{\frac{\delta_i}{\pi_i}X_i\left(Y_i - X_i^T\theta(U_i)\right) + \left(1 - \frac{\delta_i}{\pi_i}\right)\Phi_u(Z_i)\middle|U_i = u\right\} f(u)$$
$$= 0,$$

where $\pi_i = \pi(Z_i)$, and $f(u)$ is the density function of U_i. Hence, using this information, an auxiliary random vector can be defined as

$$\begin{aligned}\eta_i(\theta(u)) = &\left\{\frac{\delta_i}{\pi_i}X_i\left(Y_i - X_i^T\theta(u)\right) + \left(1 - \frac{\delta_i}{\pi_i}\right)\Phi_u(Z_i)\right\} \\ &\times K_h(U_i - u),\end{aligned}$$

where $K_h(u) = K(u/h)$, $K(\cdot)$ is a kernel function, and h is the bandwidth. For any given u, note that

$$\eta_1(\theta(u)), \cdots, \eta_n(\theta(u))$$

are independent each other, and satisfy $E\{\eta_i(\theta(u))\} = 0$ if and only if $\theta(u)$ is the true parameter. Hence using the empirical likelihood method proposed by [6], an empirical log-likelihood ratio function for $\theta(u)$ can be defined based on $\eta_i(\theta(u))$. However, $\eta_i(\theta(u))$ contains the unknown functions $\pi(z)$, $g_1(z)$ and $g_2(z)$, then it can not be used directly for the statistical inference for $\theta(u)$. A natural idea to solve this problem is to replace $\pi(z)$, $g_1(z)$ and $g_2(z)$ with the following kernel estimators respectively.

$$\hat{\pi}(z) = \frac{\sum_{i=1}^n \delta_i K_h(Z_i - z)}{\sum_{i=1}^n K_h(Z_i - z)},$$

$$\hat{g}_1(z) = \frac{\sum_{i=1}^n X_i K_h(Z_i - z)}{\sum_{i=1}^n K_h(Z_i - z)},$$

$$\hat{g}_2(z) = \frac{\sum_{i=1}^n X_i X_i^T K_h(Z_i - z)}{\sum_{i=1}^n K_h(Z_i - z)}.$$

Then, we obtain the following estimated auxiliary random vector

$$\begin{aligned}\hat{\eta}_i(\theta(u)) = &\left\{\frac{\delta_i}{\hat{\pi}_i}X_i\left(Y_i - X_i^T\theta(u)\right) + \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right)\hat{\Phi}_u(Z_i)\right\} \\ &\times K_h(U_i - u),\end{aligned} \qquad (3)$$

where $\hat{\pi}_i = \hat{\pi}(Z_i)$ and $\hat{\Phi}_u(Z_i) = Y_i\hat{g}_1(Z_i) - \hat{g}_2(Z_i)\theta(u)$. Hence, an empirical log-likelihood ratio can be given by

$$\begin{aligned}&R(\theta(u)) \\ &= -2\max\left\{\sum_{i=1}^n \log(np_i)\middle|p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i\hat{\eta}_i(\theta(u)) = 1\right\}.\end{aligned}$$

For any given u, provided that zero is inside the convex hull of the points $\left(\eta_1(\theta(u)), \cdots, \eta_n(\theta(u))\right)$, then a unique value for $R(\theta(u))$ exists. By using the Lagrange multiplier method to find the optimal p_i, then $R(\theta(u))$ can be represented as

$$R(\theta(u)) = 2\sum_{i=1}^n \log\left\{1 + \lambda^T\hat{\eta}_i(\theta(u))\right\}, \qquad (4)$$

where λ is a $p \times 1$ vector given as the solution to

$$\sum_{i=1}^n \frac{\hat{\eta}_i(\theta(u))}{1 + \lambda^T\hat{\eta}_i(\theta(u))} = 0. \qquad (5)$$

Next we will show that $R(\theta(u))$ is asymptotically chi-square distributed when $\theta(u)$ is the true parameter for given u. To derive a theory for $R(\theta(u))$, the following assumptions will be required.

Assumption 1. The bandwidth h satisfies that $nh^3 \to \infty$ and $nh^5 \to 0$.

Assumption 2. The kernel function $K(u)$ is a bounded and symmetric probability density function, and satisfies $\int u^4 K(u)du < \infty$.

Assumption 3. The density function $f(u)$ is bounded away from zero, and has continuous first derivatives. The function $\pi(z)$ has bounded partial derivatives up to the order 2 with $\inf_z \pi(z) > 0$.

Assumption 4. $\theta(u)$, $g_1(u)$ and $g_2(u)$ are twice continuously differentiable. Furthermore, we assume that $\theta_k''(u) \neq 0$, $k = 1, \cdots, p$, and $g_2(u)$ is a positive definite matrix for any given u.

Assumption5. The error ε and covariate X satisfy $\sup_u E\left(\varepsilon^4|U = u\right) < \infty$ and $\sup_u E\left(\|X\|^4|U = u\right) < \infty$, respectively, where $\|\cdot\|$ denotes the Euclidean distance.

Under these assumptions, the following theorem gives the asymptotic distribution of $R(\theta(u))$.

Theorem 1. Suppose that Assumptions 1-5 hold. For any given u, if $\theta(u)$ is the true value of the parameter, then

$$R(\theta(u)) \xrightarrow{D} \chi_p^2,$$

where "$\xrightarrow{D}$" denotes the convergence in distribution and "χ_p^2" denotes the chi-square distribution with p degrees of freedom.

By Theorem 1, the $1-\alpha$ confidence interval for $\theta(u)$ can be defined as

$$C_\alpha\left(\tilde{\theta}(u)\right) = \left\{\tilde{\theta}(u)\middle|R\left(\tilde{\theta}(u)\right) \leq \xi_\alpha\right\},$$

where ξ_α satisfies $P\left(\chi_p^2 \leq \xi_\alpha\right) = 1 - \alpha$. In addition, to implement this estimation procedure, we need to choose the bandwidth *h*. One can select *h* by optimizing some data driven criteria, such as the classical criteria CV, GCV and BIC. For the facilitation of calculation, we suggest to choose the bandwidth based on the CV criteria. More specifically, we can estimate *h* by minimizing the following cross-validation score

$$\mathrm{CV}(h)=\sum_{i=1}^{n}\delta_i\left\{Y_i-X_i^T\hat{\theta}_{[i]}(U_i)\right\}^2,$$

where $\hat{\theta}_{[i]}(u)$ is the estimator of $\theta(u)$ after deleting the *i*th subject. From our simulation experience, we found that such a choice of the bandwidth is workable.

Next we give the proof Theorem 1. The proof the Theorem 1 relies on the following lemma.

Lemma 1. Under the assumptions 1-5, we have

$$\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\xrightarrow{D}N(0,v(u)\Sigma(u)),$$

where $v(u)=f(u)\int K^2(s)ds$ and

$$\Sigma(u)=E\left\{\frac{1}{\pi(Z)}(X\varepsilon)^{\otimes 2}+\frac{\pi(Z)-1}{\pi(Z)}E(X\varepsilon|Z)^{\otimes 2}|U=u\right\},$$

Proof. From the definition of $\hat{\eta}_i(\theta(u))$ in (3), it is easy to show that

$$\begin{aligned}&\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\\&=\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}_i}X_i\left(Y_i-X_i^T\theta(u)\right)K_h(U_i-u)\\&+\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{\hat{\pi}_i}\right)\hat{\Phi}_u(Z_i)K_h(U_i-u)\\&\equiv A_1+A_2\end{aligned}\tag{6}$$

Then, similar to the proof of Theorem 4 in Wang (2009), we can prove that

$$A_1=\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(Z_i)}X_i\varepsilon_iK_h(U_i-u)+o_p(1),$$

$$A_2=\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{\pi(Z_i)}\right)E(X_i\varepsilon_i|Z_i)K_h(U_i-u)+o_p(1).$$

Hence, using the central limit theorem, we have

$$A_1\xrightarrow{D}N(0,v(u)\Sigma_1(u)),\tag{7}$$

$$A_2\xrightarrow{D}N(0,v(u)\Sigma_2(u)),\tag{8}$$

where $v(u)=f(u)\int K^2(s)ds$,

$$\Sigma_1(u)=E\left\{\frac{1}{\pi(Z)}(X\varepsilon)^{\otimes 2}|U=u\right\}$$

and $\Sigma_2(u)=E\left\{\frac{\pi(Z)-1}{\pi(Z)}E(X\varepsilon|Z)^{\otimes 2}|U=u\right\}$. Finally, this lemma follows immediately by (6) - (8).

Proof of Theorem 1. Together with the proof of Lemma 1 and using the same argument as are used in the proof of Lemma 1 in [7], we can show that

$$\max_{1\le i\le n}\left\|\hat{\eta}_i(\theta(u))\right\|=o_p\left((nh)^{1/2}\right).\tag{9}$$

Similar to the proof of (2.14) in [6], we can prove that

$$\|\lambda\|=O_p\left((nh)^{-1/2}\right).\tag{10}$$

Then, invoking (9) and (10), and applying the Taylor expansion to (4), it is easy to show that

$$\begin{aligned}&R(\theta(u))\\&=2\sum_{i=1}^{n}\left\{\lambda^T\hat{\eta}_i(\theta(u))-\left[\lambda^T\hat{\eta}_i(\theta(u))\right]^2/2\right\}+o_p(1).\end{aligned}\tag{11}$$

Furthermore, from (5) and invoking (9) and (10), we can prove that

$$\begin{aligned}\lambda=&\left(\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\hat{\eta}_i^T(\theta(u))\right)^{-1}\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\\&+o_p\left((nh)^{-1/2}\right),\end{aligned}\tag{12}$$

$$\sum_{i=1}^{n}\left[\lambda^T\hat{\eta}_i(\theta(u))\right]^2=\sum_{i=1}^{n}\lambda^T\hat{\eta}_i(\theta(u))+o_p(1).\tag{13}$$

Using (11)-(13), we obtain that

$$\begin{aligned}R(\theta(u))=&\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\right)^T\hat{\Sigma}(u)^{-1}\\&\cdot\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\right).\end{aligned}\tag{14}$$

where $\hat{\Sigma}(u)=\frac{1}{nh}\sum_{i=1}^{n}\hat{\eta}_i(\theta(u))\hat{\eta}_i^T(\theta(u))$. Invoking the proof of the Lemma 1 and using the law of large numbers, we obtain that

$$\hat{\Sigma}(u)\xrightarrow{P}v(u)\Sigma(u).$$

This together with (14) and Lemma 1 yields Theorem 1.

3. Simulation Studies

In this section, some Monte Carlo simulations are conducted to evaluate the finite sample performance of the proposed empirical likelihood method. The data are generated from the following model

$$Y=X\theta(u)+\varepsilon,$$

where $\theta(u)=\sin(2\pi u)$, the covariates U and X are generated according to $U\sim U(0,1)$ and $X\sim N(0,1)$, respectively. The response Y is generated according the model with $\varepsilon \sim N(0,0.5)$. In the following simulation procedure, we choose the following two missing data mechanism:

Case1:

$$\pi(y,u) = \exp(1+0.5y+0.45u)/\left(0.5+\exp(1+0.5y+0.45u)\right),$$

Case 2:

$$\pi(y,u) = \exp(1+0.5y+0.45u)/\left(1+\exp(1+0.5y+0.45u)\right)$$

The average missing rates of these two cases are 0.15 and 0.25 respectively. For each case, we take 1000 simulation runs. In addition, the sample size is taken as n = 200.

For comparison, we consider two methods for constructing the confidence intervals: the imputed estimation method (IEL) proposed by this paper, and the naïve empirical likelihood method (NEL). The latter is neglecting the incomplete data information, and constructing the confidence intervals for the regression coefficients only based on the complete data. The averages of the confidence intervals with the nominal level $1-\alpha = 95\%$, computed with 1000 simulation runs, are summarized in **Figures 1** and **2**. **Figure 1** is the simulation results under the missing mechanism Case 1, and **Figure 2** is the simulation results under the missing mechanism Case 2, where the dashed curves mean the results obtained by IEL method, the dotted curves mean the results obtained by NEL method, and the solid curve represents the real curve of $\theta(u)$.

From **Figures 1** and **2**, we can make the following observations:

(i) The confidence intervals based on the IEL method outperform those based on the NEL method, because lengths of the confidence intervals obtained by the IEL method are shorter than those obtained by the NEL method.

(ii) The performances of the confidence intervals based on the IEL method are similar for all levels of missing mechanisms. This implies that the imputed empirical likelihood procedure can attenuate the effect of missing data.

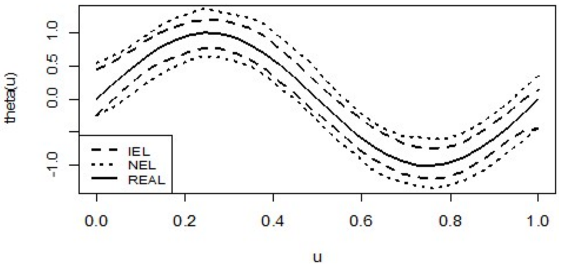

Figure 1. The 95% confidence intervals of $\theta(u)$ under the missing mechanism Case 1 based on IEL method (dashed curve) and NEL method (dotted curve).

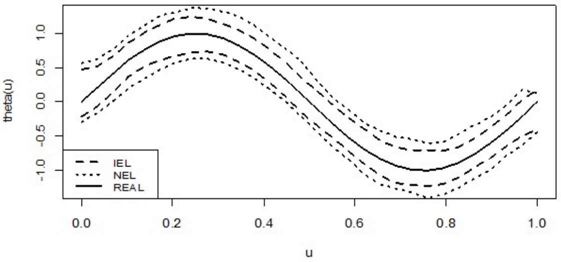

Figure 2. The 95% confidence intervals of $\theta(u)$ under the missing mechanism Case 2 based on IEL method (dashed curve) and NEL method (dotted curve).

4. Conclusions and Discussions

We have proposed an imputed empirical likelihood procedure for varying coefficient models when some covariates are missing. The proposed method can attenuate the effect of missing data efficiently, and extends the imputation-based estimation method to the varying coefficient models with missing covariates. Simulation studies indicated that the proposed method was very effective in attenuating the effect of missing data and constructing the confidence intervals for the coefficient functions.

In this paper, although we assume that all components of the covariate are subject to missing, it is not essential. The proposed estimation method can easily extend the case that only some components of the covariate are measured with missing. In addition, one useful extension of the varying coefficient model is the varying coefficient partially linear model. For such model, Zhao and Xue [8] considered the statistical inferences for regression coefficients when the response with missing. Then, another interesting topic of further research is investigating the inferences for such varying coefficient partially linear models with missing covariates.

REFERENCES

[1] Q. H. Wang and J. N. K. Rao, “Empirical Likelihood for Linear Models under Imputation for Missing Responses,” The *Canadian Journal of Statistics,* Vol. 29, No. 4，2001, pp. 597-608 . doi：10.2307/3316009

[2] L. G. Xue, “Empirical Likelihood for Linear Models with Missing Responses,” *Journal of Multivariate Analysis,* Vol. 100, No. 7,2009, pp. 1353-1366. doi：10.1016/j.jmva.2008.12.009

[3] Q. H. Wang, “Statistical Estimation in Partial Linear Models with Covariate Data Missing at Random,” *Annals of the Institute of Statistical Mathematics*, Vol. 61, No. 1, 2009, pp. 47-84. doi:10.1007/s10463-007-0137-1

[4] H. Liang, S. J. Wang, J. M. Robbins and R. J. Carroll, “Estimation in Partially Linear Models with Missing

Covariates," *Journal of the American Statistical Association*, Vol. 99, No. 466, 2004, pp. 357-367. doi：10.1198/016214504000000421

[5] H. Wong, S. J. Guo, M. Chen and W.C. IP, "On Locally Weighted Estimation and Hypothesis Testing of Varying Coefficient Models with Missing Covariates," *Journal of Statistical Planning and Inference*, Vol. 139, No. 9, No. 1, 2009, pp. 2933-2951. doi：10.1016/j.jspi.2009.01.016

[6] A. B. Owen, "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, Vol. 18, No. 1, 1990, pp. 90-120. doi：10.1214/aos/1176347494

[7] L. G. Xue and L. X. Zhu, "Empirical Likelihood for a Varying Coefficient Model with Longitudinal Data," *Journal of the American Statistical Association*, Vol. 102, No. 478, 2007, pp. 642-654. doi：10.1198/016214507000000293

[8] P. X. Zhao and L. G. Xue, "Variable Selection for Semiparametric Varying Coefficient Partially Linear Models with Missing Response at Random," *Acta Mathematica Sinica*, *English Series*, Vol. 27, No. 11, 2011, pp. 2205-2216. doi：10.1007/s10114-011-9200-1