

Functional Coefficient Estimation with Both Categorical and Continuous Data

Liangjun Su,^a Ye Chen,^b Aman Ullah^c

^aSchool of Economics, Singapore Management University, Singapore

^bDepartment of Economics, Princeton University, NJ, USA

^cDepartment of Economics, University of California at Riverside, CA, USA

June 18, 2009

Abstract

We propose a local linear functional coefficient estimator that admits a mix of discrete and continuous data for stationary time series. Under weak conditions our estimator is asymptotically normally distributed. A small set of simulation studies is carried out to illustrate the finite sample performance of our estimator. As an application, we estimate a wage determination function that explicitly allows the return to education to depend on other variables. We find evidence of the complex interacting patterns among the regressors in the wage equation, such as increasing returns to education when experience is very low, high return to education for workers with several years of experience, and diminishing returns to education when experience is high. Compared with the commonly used parametric and semi-parametric methods, our estimator performs better in both goodness-of-fit and in yielding economically interesting interpretation.

Key words: Discrete variables, Functional coefficient estimation, Local linear estimation, Least squares cross validation.

JEL Classification: C13, C14.

1 Introduction

In this paper we extend the work of Racine and Li (2004) to estimating functional coefficient models with both continuous and categorical data:

$$Y = \sum_{j=1}^d a_j(U) X_j + \varepsilon, \quad (1.1)$$

where ε is the disturbance term, X_j is a scalar random variable, U is a $(p+q) \times 1$ random vector, and $a_j(\cdot)$, $j = 1, \dots, d$, are unknown smooth functions. As Cai, Fan and Yao (2000) remark, the idea for this kind of model is not new, but the potential of this modeling techniques had not been fully explored until the seminal work of Cleveland et al. (1992), Chen and Tsay (1993), and Hastie and Tibshirani (1993), in which nonparametric techniques were proposed to estimate the unknown functions $a_j(\cdot)$. An important feature of these early works is to assume that the random variable U is continuous, which limits the model's potential applications.

Drawing upon the work of Aitchison and Aitken (1976), Racine and Li (2004) propose a novel approach to estimate nonparametric regression mean functions with both categorical and continuous data in the i.i.d. setup. They apply their new estimation method to some publicly available data and demonstrate the superb performance of their estimators in comparison with some traditional ones.

In this paper, we consider extending the work of Racine and Li (2004) to the estimation of the functional coefficient model (1.1) when U contains both continuous and categorical variables. This is important since categorical variables may be present in the functional coefficients. For example, in the study of the output functions for individual firms, firms that belong to different industries may exhibit different output elasticities with respect to labor and capital. So we should allow the categorical variable 'industry' to enter U . We will demonstrate that this modelling strategy outperforms the traditional dummy-variable approach widely used in the literature. For the same reason, Li and Racine (2008b) consider a local constant estimation of model (1.1) by assuming the data are identically and independently distributed (i.i.d.).

Another distinguishing feature of our approach is that we allow for weak data dependence. One of the key applications of nonparametric function estimation is the construction of prediction intervals for stationary time series. The i.i.d. setup of Racine and Li (2004) and Li and Racine (2008b) cannot meet this purpose.

To demonstrate the usefulness of our proposed estimator in empirical applications, we estimate a wage determination equation based on recent CPS data. While in the literature of labor economics, the return to education has already been extensively investigated from various aspects, in this paper, we explicitly allow the return to education to be dependent on other variables, both continuous and discrete, including experience, gender, age, industry and so forth. Our findings are clearly against the parametric functional form assumption of the most widely used linear separable Mincerian equation, and the return to education does vary substantially with the other regressors. Therefore, our model can help to uncover economically interesting interacting effects among the regressors, and so should have high potential for applications.

The paper is structured as follows. In Section 2 we introduce our functional coefficient estimators and their asymptotic properties. We conduct a small set of Monte Carlo studies to check the relative performance of the proposed estimator in Section 3. Section 4 provides empirical data analysis. Final remarks are contained in Section 5. All technical details are relegated to the Appendix.

2 Functional Coefficient Estimation with Mixed Data

2.1 Local linear estimator

In this paper, we study estimation of model (1.1) when U is comprised of a mix of discrete and continuous variables. Let $\{(Y_i, X_i, U_i), i = 1, 2, \dots\}$ be jointly strictly stationary processes, where (Y_i, X_i, U_i) has the same distribution as (Y, X, U) . Let $U_i = (U_i^c, U_i^d)'$, where U_i^c and U_i^d denote a $p \times 1$ vector of continuous regressors and a $q \times 1$ vector of discrete regressors, respectively, $p \geq 1$, and $q \geq 1$. Like Racine and Li (2004), we will use U_{it}^d to denote the t th component of U_i^d , and assume that U_{it}^d can take $c_t \geq 2$ different values, i.e., $U_{it}^d \in \{0, 1, \dots, c_t - 1\}$ for $t = 1, \dots, q$. Denote $u = (u^c, u^d) \in \mathbb{R}^p \times \mathcal{D}$. We use $f_u(u) = f(u^c, u^d)$ to denote the joint density function of (U_i^c, U_i^d) and $\mathcal{D} = \prod_{t=1}^q \{0, 1, \dots, c_t - 1\}$ to denote the range assumed by U_i^d . With a little abuse of notation, we also use $\{(Y_i, X_i, U_i), i = 1, \dots, n\}$ to denote the data.

To define the kernel weight function, we focus on the case for which there is no natural ordering in U_i^d . Define

$$l(U_{it}^d, u_t^d, \lambda_t) = \begin{cases} 1 & \text{if } U_{it}^d = u_t^d, \\ \lambda_t & \text{if } U_{it}^d \neq u_t^d, \end{cases} \quad (2.1)$$

where λ_t is a bandwidth that lies on the interval $[0, 1]$. Clearly, when $\lambda_t = 0$, $l(U_{it}^d, u_t^d, 0)$ becomes an indicator function, and $\lambda_t = 1$, $l(U_{it}^d, u_t^d, 1)$ becomes an uniform weight function. We define the product kernel for the discrete random variables by

$$L(U_i^d, u^d, \lambda) = \prod_{t=1}^q l(U_{it}^d, u_t^d, \lambda_t). \quad (2.2)$$

For the continuous random variables, we use $w(\cdot)$ to denote a univariate kernel function and define the product kernel function by $W_{h,iu} = \prod_{t=1}^p w((U_{it}^c - u_t^c)/h_t)$, where $h = (h_1, \dots, h_p)$ denotes the smoothing parameters and U_{it}^c (u_t^c) is the t th component of U_i^c (u^c). We then define the kernel weight function K_{iu} by

$$K_{iu} = L_{\lambda,iu} W_{h,iu} \quad (2.3)$$

where $L_{\lambda,iu} = L(U_i^d, u^d, \lambda)$.

We now estimate the unknown functional coefficient functions in model (1.1) by using a local linear regression technique. Suppose that $a_j(\cdot)$ assumes a second order derivative. Denote by $\dot{a}_j(u) = \partial a_j(u) / \partial u^c$ the $p \times 1$ first order derivative of $a_j(u)$ with respect to its continuous-valued argument u^c . Denote by $\ddot{a}_j(u) = \partial^2 a_j(u) / (\partial u^c \partial u^c)$ the $p \times p$ second order derivative matrix of $a_j(u)$ with respect to u^c . We use $a_{j,ss}(u)$ to denote the s th diagonal element of $\ddot{a}_j(u)$.

For any given u and \tilde{u} in a neighborhood of u , it follows from a first order Taylor expansion that

$$a_j(\tilde{u}) \approx a_j(u) + \dot{a}_j(u)'(\tilde{u}^c - u^c), \quad (2.4)$$

for u^c in a neighborhood of \tilde{u}^c and $\tilde{u}^d = u^d$. To estimate $\{a_j(u)\}$ (and $\{\dot{a}_j(u)\}$), we choose $\{a_j\}$ and $\{b_j\}$ to minimize

$$\sum_{i=1}^n \left[Y_i - \sum_{j=1}^d \{a_j + b'_j(U_i - u)\} X_{ij} \right]^2 K_{iu}. \quad (2.5)$$

Let $\{(\hat{a}_j, \hat{b}_j)\}$ be the local linear estimator. Then the local linear regression estimator for the functional coefficient is given by

$$\hat{a}_j(u) = \hat{a}_j, \quad j = 1, \dots, d. \quad (2.6)$$

The local linear regression estimator for the functional coefficient can be easily obtained. To do so, let $\mathbf{e}_{j,d(1+p)}$ be the $d(1+p) \times 1$ unit vector of with 1 at the j th position and 0 elsewhere. Let $\tilde{\mathbf{X}}$ denote an $n \times d(1+p)$ matrix with

$$\tilde{X}_i = (X'_i, X'_i \otimes (U_i - u)')$$

as its i th row. Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$. Set $\mathbf{W} = \text{diag}\{K_{1u}, \dots, K_{nu}\}$. Then (2.5) can be written as

$$(\mathbf{Y} - \tilde{\mathbf{X}}\theta)' \mathbf{W} (\mathbf{Y} - \tilde{\mathbf{X}}\theta),$$

where $\theta = (a_1, \dots, a_d, b'_1, \dots, b'_d)'$. So the local linear estimator is simply

$$\hat{\theta} = \hat{\theta}(u) = (\tilde{\mathbf{X}}' \mathbf{W} \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}' \mathbf{W} \mathbf{Y}, \quad (2.7)$$

which entails that

$$\hat{a}_j = \hat{a}_j(u) = \mathbf{e}'_{j,d(1+p)} \hat{\theta}, \quad j = 1, \dots, d. \quad (2.8)$$

Let $\theta(u) = (a_1(u), \dots, a_d(u), \dot{a}_1(u)', \dots, \dot{a}_d(u)')'$. We will study the asymptotic properties of $\hat{\theta}(u)$.

2.2 Assumptions

To facilitate the presentation, let $\Omega(u) = E(X_i X'_i | U_i = u)$, $\sigma^2(u, x) = E[\varepsilon_i^2 | U_i = u, X_i = x]$, $\Omega^*(u) = E[X_i X'_i \sigma^2(U_i, X_i) | U_i = u]$. Let $f(u, x)$ denote the joint density of (U_i, X_i) and $f_u(u)$ be the marginal density of U_i . Also, let $f_{u|x}(u|x)$ be the conditional density of U_i given $X_i = x$. Let $f_i(u, \tilde{u}|x, \tilde{x})$ be the conditional density of (U_1, U_i) given $(X_1, X_i) = (x, \tilde{x})$.

We now list the assumptions that will be used to establish the asymptotic distribution of our estimator.

Assumption

A1. (i) The process $\{(Y_i, U_i, X_i), i \geq 1\}$ is a strictly stationary α -mixing process with coefficients $\alpha(n)$ satisfying $\sum_{j \geq 1} j^c [\alpha(j)]^{\gamma/(2+\gamma)} < \infty$ for some $\gamma > 0$ and $c > \gamma/(2+\gamma)$.

(ii) $f_{u|x}(u|x) \leq M < \infty$ and $f_i(u, \tilde{u}|x, \tilde{x}) \leq M < \infty$ for all $i \geq 2$ and $u, \tilde{u}, x, \tilde{x}$.

(iii) $\Omega^*(u)$ and $\Omega(u)$ are positive definite.

(iv) The functions $f_u(\cdot, u^d)$, $\sigma^2(\cdot, u^d, x)$, $\Omega(\cdot, u^d)$, and $\Omega^*(\cdot, u^d)$ are continuous for all $u^d \in \mathcal{D}$, and $f_u(u) > 0$.

(v) $a_j(\cdot, u^d)$ has continuous second derivatives for all $u^d \in \mathcal{D}$.

(vi) $E \|X\|^{2(2+\gamma)} < \infty$, where $\|\cdot\|$ is the Euclidean norm and γ is given in (i).

(vii) $E [Y_1^2 + Y_i^2 | (U_1, X_1) = (u, x); (U_i, X_i) = (\tilde{u}, \tilde{x})] \leq M < \infty$.

(viii) There exists $\delta > 2 + \gamma$ such that $E [|Y_1|^\delta | (U_1, X_1) = (\tilde{u}, x)] \leq M < \infty$ for all $x \in \mathbb{R}^d$ and all \tilde{u} in the neighborhood of u . $\alpha(j) = O(j^{-\kappa})$, where $\kappa \geq (2 + \gamma)\delta / \{2(\delta - 2 - \gamma)\}$.

(ix) There exists a sequence of positive integers s_n such that $s_n \rightarrow \infty$, $s_n = o((nh_1 \dots h_p)^{1/2})$, and $n^{1/2} (h_1 \dots h_p)^{-1/2} \alpha(s_n) \rightarrow 0$.

A2. The kernel function $w(\cdot)$ is a density function that is symmetric, bounded, and compactly supported.

A3. As $n \rightarrow \infty$, the bandwidth sequences $h_s \rightarrow 0$ for $s = 1, \dots, p$, $\lambda_s \rightarrow 0$ for $s = 1, \dots, q$, and (i) $nh_1 \dots h_p \rightarrow \infty$, (ii) $(nh_1 \dots h_p)^{1/2} (\|h\|^2 + \|\lambda\|) = O(1)$.

Assumptions A1-A2 are similar to Conditions A and B in Cai, Fan and Yao (2000) except that we consider mixed regressors. Assumption A1(i) is standard in the nonparametric regression for time series. See, for example, Cai, Fan and Yao (2000), and Cai and Ould-Saïd (2003). It is satisfied by many well-known processes such as linear stationary ARMA processes and a large class of processes implied by numerous nonlinear models, including bilinear, nonlinear autoregressive (NLAR), and ARCH-type models (see Fan and Li, 1999). As Hall et al. (1999) and Cai and Ould-Saïd (2003) remark, the requirement in Assumption A2 that $w(\cdot)$ is compactly supported can be removed at the cost of lengthier arguments used in the proofs, and in particular, Gaussian kernel is allowed.

Assumption A3 is standard for nonparametric regression with mixed data (see Li and Racine, 2008a).

2.3 Asymptotic theory for the local linear estimator

To introduce our main results, let $\mu_{s,t} = \int_{\mathbb{R}} v^s w(v)^t dv$, $s, t = 0, 1, 2$. Define two $d(1+p) \times d(1+p)$ diagonal matrices $S = S(u)$ and $\Gamma = \Gamma(u)$ by

$$S = f_u(u) \begin{pmatrix} \Omega(u) & 0'_{dp \times d} \\ 0_{dp \times d} & \mu_{2,1} \Omega(u) \otimes I_p \end{pmatrix}, \quad \Gamma = f_u(u) \begin{pmatrix} \mu_{0,2}^p \Omega^*(u) & 0'_{dp \times d} \\ 0_{dp \times d} & \mu_{2,2} \Omega^*(u) \otimes I_p \end{pmatrix},$$

where $0_{l \times k}$ is an $l \times k$ matrix of zeros, I_p is the $p \times p$ identity matrix, and \otimes is the Kronecker product. For any $p \times 1$ vectors $c = (c_1, \dots, c_p)'$ and $d = (d_1, \dots, d_p)'$, let $c \odot d \equiv (c_1 d_1, \dots, c_p d_p)'$.

To describe the leading bias term associated with the discrete random variables, we define

$$I_s(u^d, \tilde{u}^d) = \mathbf{1}(u_s^d \neq \tilde{u}_s^d) \prod_{t \neq s}^q \mathbf{1}(u_t^d = \tilde{u}_t^d),$$

where $\mathbf{1}(\cdot)$ is the usual indicator function. That is, $I_s(u^d, \tilde{u}^d)$ is one if and only u^d and \tilde{u}^d differ only

in the s th component and is zero otherwise. Let

$$b(h, \lambda) = H \left\{ \begin{aligned} & \left(\begin{array}{c} \frac{1}{2} \mu_{2,1} f_u(u) \Omega(u) A \\ 0_{d_p \times 1} \end{array} \right) \\ & + \sum_{\tilde{u}^d \in \mathcal{D}} \sum_{s=1}^q \lambda_s I_s(u^d, \tilde{u}^d) f_u(u^c, \tilde{u}^d) \left(\begin{array}{c} \Omega(u^c, \tilde{u}^d) (\mathbf{a}(u^c, \tilde{u}^d) - \mathbf{a}(u)) \\ -\mu_{2,1} (\Omega(u^c, \tilde{u}^d) \otimes I_p) \mathbf{b}(u) \end{array} \right) \end{aligned} \right\}, \quad (2.9)$$

where $H = \sqrt{nh_1 \dots h_p}$, $A = (\sum_{s=1}^p h_s^2 a_{1,ss}(u), \dots, \sum_{s=1}^p h_s^2 a_{d,ss}(u))'$, $\mathbf{a}(u) = (a_1(u), \dots, a_d(u))'$, and $\mathbf{b}(u) = (\dot{a}_1(u)', \dots, \dot{a}_d(u)')$. Define

$$\begin{aligned} B_{j,1s}(u) &= \frac{1}{2} \mu_{2,1} a_{j,ss}(u), \text{ and} \\ B_{j,2s}(u) &= f_u(u)^{-1} \mathbf{e}'_{j,d} \Omega^{-1}(u) \sum_{\tilde{u}^d \in \mathcal{D}} I_s(u^d, \tilde{u}^d) f(u^c, \tilde{u}^d) \Omega(u^c, \tilde{u}^d) [\mathbf{a}(u^c, \tilde{u}^d) - \mathbf{a}(u)]. \end{aligned}$$

Now we state our main theorem.

Theorem 2.1 *Assume that Assumptions A1-A3 hold. Then for each u that is an interior point*

$$HH_1 \left(\hat{\theta}(u) - \theta(u) \right) - S^{-1} b(h, \lambda) \xrightarrow{d} N(0, S^{-1} \Gamma S^{-1}).$$

where $H_1 = \text{diag}(1, \dots, 1, h', \dots, h')$ is a $d(p+1) \times 1$ diagonal matrix with d diagonal elements of 1 and d diagonal elements of h . In particular, for $j = 1, \dots, d$,

$$\begin{aligned} & \sqrt{nh_1 \dots h_p} \left(\hat{a}_j(u) - a_j(u) - \sum_{s=1}^p h_s^2 B_{j,1s}(u) - \sum_{s=1}^q \lambda_s B_{j,2s}(u) \right) \\ & \xrightarrow{d} N \left(0, \frac{\mu_{0,2}^p \mathbf{e}'_{j,d} \Omega^{-1}(u) \Omega^*(u) \Omega^{-1}(u) \mathbf{e}_{j,d}}{f_u(u)} \right). \end{aligned}$$

Remark 1. Noting that S and Γ are both block diagonal matrices, we have asymptotic independence between the estimator of $\mathbf{a}(u)$ and that of $\mathbf{b}(u)$. Under Assumption A3, the asymptotic bias of \hat{a}_j is comprised of two components, $\sum_{s=1}^p h_s^2 B_{j,1s}(u)$ and $\sum_{s=1}^q \lambda_s B_{j,2s}(u)$, which are associated with the continuous and discrete variables in U_i , respectively. For statistical inference, one needs to estimate $f_u(u)$, $\Omega(u)$ and $\Omega^*(u)$. The procedure is standard and thus is omitted.

Remark 2. It is well known that the two main advantages of a local linear estimate over a local constant estimate are the simpler structure of asymptotic bias and the automatic boundary bias correction mechanism for the local linear estimate (see Fan and Gijbels (1996)). Our local linear estimator has the same asymptotic variance as the local constant estimator of Li and Racine (2008b). But the two estimators are different in bias. In our notation, the asymptotic bias (Abias) of Li and Racine's local constant estimator $\hat{a}_j^{(lc)}(u)$ of $a_j(u)$ is given by

$$\text{Abias}(\hat{a}_j^{(lc)}(u)) = \sum_{s=1}^p h_s^2 B_{j,1s}^{(lc)}(u) - \sum_{s=1}^q \lambda_s B_{j,2s}^{(lc)}(u),$$

where

$$\begin{aligned} B_{j,1s}^{(lc)}(u) &= \mu_{2,1} \left\{ \mathbf{e}'_{j,d} f_u(u)^{-1} \Omega^{-1}(u) [f_u(u) \Omega_s(u) + \Omega(u) f_{u,s}(u)] \mathbf{a}_s(u) + \frac{1}{2} a_{j,ss}(u) \right\}, \\ B_{j,2s}^{(lc)}(u) &= B_{j,2s}(u), \end{aligned}$$

$\Omega_s(u)$ denotes the first order partial derivative of $\Omega(u^c, u^d)$ with respect to the s th element in u^c , and $f_{u,s}(u)$ and $\mathbf{a}_s(u)$ are similarly defined. Clearly, the continuous element in $u = (u^c, u^d)$ causes the difference in the asymptotic biases of the two types of estimators.

To compare boundary behavior of the two estimators, we focus on the simplest case where there is only one continuous variable in $U_i = (U_i^c, U_i^d)'$, i.e., U_i^c is a scalar random variable and $p = 1$. Without loss of generality, we assume that the support of U_i^c is $[0, 1]$. In this case, we denote the bandwidth simply as $h \equiv h(n)$ and consider the left boundary point $u^c = \nu h$, where ν is a finite positive constant. Following the literature, we assume that $f_u(0, u^d) \equiv \lim_{u^c \downarrow 0} f_u(u^c, u^d)$ exists and is strictly positive for all $u^d \in \mathcal{D}$. Define

$$S_\nu = \begin{pmatrix} \iota_{\nu 0} & \iota_{\nu 1} \\ \iota_{\nu 1} & \iota_{\nu 2} \end{pmatrix}, \text{ and } \Gamma_\nu = \begin{pmatrix} \kappa_{\nu 0} & \kappa_{\nu 1} \\ \kappa_{\nu 1} & \kappa_{\nu 2} \end{pmatrix}, \quad (2.10)$$

where $\iota_{\nu j} = \int_{-\nu}^{\infty} z^j w(z) dz$, and $\kappa_{\nu j} = \int_{-\nu}^{\infty} z^j w(z)^2 dz$ for $j = 0, 1$, and 2 . Define

$$S(0, u^d; \nu) = S_\nu \otimes \Omega(0, u^d) f_u(0, u^d), \text{ and } \Gamma(0, u^d; \nu) = \Gamma_\nu \otimes \Omega^*(0, u^d) f_u(0, u^d).$$

Define

$$\begin{aligned} b(h, \lambda; \nu) &= H \left\{ \frac{1}{2} \begin{pmatrix} \Omega(0, u^d) \bar{A}(0, u^d) \iota_{\nu 2} \\ \Omega(0, u^d) \bar{A}(0, u^d) \iota_{\nu 3} \end{pmatrix} f_u(0, u^d) \right. \\ &\quad \left. + \sum_{\tilde{u}^d \in \mathcal{D}} \sum_{s=1}^q \lambda_s I_s(u^d, \tilde{u}^d) f_u(0, \tilde{u}^d) \begin{pmatrix} \Omega(0, \tilde{u}^d) \{ \iota_{\nu 0} [\mathbf{a}(0, \tilde{u}^d) - \mathbf{a}(0, u^d)] - \iota_{\nu 1} \mathbf{b}(0, u^d) \} \\ \Omega(0, \tilde{u}^d) \{ \iota_{\nu 1} [\mathbf{a}(0, \tilde{u}^d) - \mathbf{a}(0, u^d)] - \iota_{\nu 2} \mathbf{b}(0, u^d) \} \end{pmatrix} \right\}, \end{aligned}$$

where $\iota_{\nu 3} = \int_{-\nu}^{\infty} z^3 w(z) dz$,

$$\bar{A}(0, u^d) = (h^2 a_1''(0, u^d), \dots, h^2 a_d''(0, u^d))', \quad (2.11)$$

and $a_s''(0, u^d)$ is the second order derivative of $a_s(u^c, u^d)$ with respect to u^c evaluated at 0 . The following corollary summarizes the asymptotic properties of $\hat{\theta}(u) = \hat{\theta}(u^c, u^d)$ for the case where $u^c = \nu h$.

Corollary 2.2 *Assume that Assumptions A1-A3 hold. If $p = 1$ and the support of U_i^c is $[0, 1]$, then for any $u = (u^c, u^d)$ with $u^c = \nu h$, we have*

$$HH_1 \left(\hat{\theta}(u) - \theta(u) \right) - S(0, u^d; \nu)^{-1} b(h, \lambda; \nu) \xrightarrow{d} N \left(0, S(0, u^d; \nu)^{-1} \Gamma(0, u^d; \nu) S(0, u^d; \nu)^{-1} \right).$$

Remark 3. Clearly, for our local linear estimators the biases for the boundary points have the same order as those for the interior points. But the estimators of $\mathbf{a}(u)$ and $\mathbf{b}(u)$ are generally not

asymptotically independent any more because neither $S(0, u^d; \nu)$ nor $\Gamma(0, u^d; \nu)$ is block diagonal. As a result, the asymptotic bias and variance formulae of $\hat{a}_j(u)$ are not as simple as those in Theorem 2.1. Li and Racine (2008b) did not study the boundary behavior of the local constant estimator. Nevertheless, following the arguments used in the proof of the above corollary, we can readily show that their estimator has the same asymptotic variance as ours for boundary points but totally different bias formula. In our notation, the asymptotic bias (Abias) of Li and Racine's local constant estimator $\hat{\mathbf{a}}^{(lc)}(u^c, u^d)$ of $\mathbf{a}(u^c, u^d)$ with $u^c = \nu h$ (after being scaled by H) is given by

$$\text{Abias}(\hat{\mathbf{a}}^{(lc)}(u^c, u^d)) = S^{(lc)}(0, u^d; \nu)^{-1} b^{(lc)}(h, \lambda; \nu),$$

where $S^{(lc)}(0, u^d; \nu) = \iota_{\nu 0} \Omega(0, u^d) f_u(0, u^d)$,

$$\begin{aligned} b^{(lc)}(h, \lambda; \nu) &= H \left\{ f_u(0, u^d) \Omega(0, u^d) \bar{A}^{(lc)}(0, u^d) \iota_{\nu 1} \right. \\ &\quad \left. + \sum_{\tilde{u}^d \in \mathcal{D}} \sum_{s=1}^q \lambda_s I_s(u^d, \tilde{u}^d) f_u(0, \tilde{u}^d) \Omega(0, \tilde{u}^d) [\mathbf{a}(0, \tilde{u}^d) - \mathbf{a}(0, u^d)] \right\}, \end{aligned}$$

and

$$\bar{A}^{(lc)}(0, u^d) = (h \dot{a}_1(0, u^d), \dots, h \dot{a}_d(0, u^d))'. \quad (2.12)$$

That is, the contribution of the continuous variable U_i^c to the asymptotic bias of the boundary estimator is of order $O(h)$, which is different from the order $O(h^2)$ for interior points. This is a reflection of the main disadvantage of local constant estimators over the local linear estimators.

2.4 Selection of smoothing parameters

In this subsection we focus on how to choose the smoothing parameters to obtain the estimate \hat{a}_j . It is well known that the choice of smoothing parameters is crucial in nonparametric kernel estimation.

Theorem 2.1 implies that the leading term for the mean squared error (MSE) of \hat{a}_j is

$$MSE(\hat{a}_j) = \left[\sum_{s=1}^p h_s^2 B_{j,1s}(u) + \sum_{s=1}^q \lambda_s B_{j,2s}(u) \right]^2 + \frac{1}{nh_1 \dots h_p} \frac{\mu_{0,2}^p \mathbf{e}'_{j,d} \Omega^{-1}(u) \Omega^*(u) \Omega^{-1}(u) \mathbf{e}_{j,d}}{f_u(u)}.$$

By symmetry, all h_j should have the same order and all λ_s should also have the same order but with $\lambda_s \sim h_j^2$. By an argument similar to Li and Racine (2008a), it is easy to obtain the optimal rate of bandwidth in terms of minimizing a weighted integrated version of $MSE(\hat{a}_j)$. To be concrete, we should choose

$$h_j \sim n^{-1/(4+p)} \text{ and } \lambda_j \sim n^{-2/(4+p)}.$$

Nevertheless, the exact formula for the optimal smoothing parameters is difficult to obtain except for the simplest cases (e.g., $p = 1$ and $q = 1$). This also suggests that it is infeasible to use the plug-in bandwidth in applied setting since the plug-in method would first require the formula for each smoothing parameter and then pilot estimates for some unknown functions in the formula.

In practice, the key in estimating the functional coefficient model is the selection of bandwidth. We propose to use least squares cross-validation (LSCV) to choose the smoothing parameters. We choose (h, λ) to minimize the following LSCV criterion function

$$CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \sum_{j=1}^d \hat{a}_j^{(-i)}(U_i) X_{ij} \right)^2 M(U_i), \quad (2.13)$$

where $\hat{a}_j^{(-i)}(U_i)$ is the leave-one-out functional coefficient estimator of $a_j(U_i)$ and $M(U_i)$ is a weight function that serves to avoid division by zero and perform trimming in areas of sparse support. In the following numerical study we will set $M(U_i) = \prod_{j=1}^p \mathbf{1}(|U_{ij}^c - \bar{U}_j^c| \leq 2s_{U_j^c})$, where $\mathbf{1}(\cdot)$ is the usual indicator function, and \bar{U}_j^c and $s_{U_j^c}$ denote the sample mean and standard deviation of $\{U_{ij}^c, 1 \leq i \leq n\}$, respectively. In practice, we can use grid search for (h, λ) when the dimensions of U^c and U^d are both small. Alternatively, one can apply the minimization function built in various software; but multiple starting values are recommended to reduce the chance of local solutions. In the following simulation study with $p = 1$ and $q = 2$, we try to save time in computation and use the latter method with only one starting value set according to the rule of thumb: $h_0 = S_{U^c} n^{-1/5}$, $\lambda_j = 0.5 S_{U^c} n^{-2/5}$ for $j = 1, 2$, where S_{U^c} is the standard deviation of the scalar random variable U_i^c . The performance of our nonparametric estimator is already reasonably well with this simple method.

Nevertheless, if the number of observations in application is large, it is extremely costly to apply the above LSCV method directly on all the observations. So we now propose an alternative way to do the LSCV. But the theoretical justification of this novel approach is beyond the scope of this paper. Let n denote the number of observations in the dataset, which could be as large as 17446 in our empirical applications. When there is only one continuous variable in U (i.e., U^c is a scalar and $p = 1$), we propose the following approach to obtain the data-driven bandwidth:

Step 1. For $b = 1, 2, \dots, B$, sample $m (\ll n)$ observations randomly from the dataset.

Step 2. Set $h = c S_{U^c} m^{-1/5}$ and $\lambda_j = c_j S_{U^c} m^{-2/5}$ for each and $j = 1, \dots, q$, where c and c_j take values on $[0.2, 4]$ with increments 0.2 and with the constraint $\lambda_j \leq 1$ satisfied, and S_{U^c} is the standard deviation of U_i^c based on the m observations in Step 1. Find the values of c and c_j that minimize the LSCV criterion function. Denote them as $c^{(b)}$ and $c_j^{(b)}$ for the b th resample.

Step 3. Calculate $\bar{c} = B^{-1} \sum_{b=1}^B c^{(b)}$ and $\bar{c}_j = B^{-1} \sum_{b=1}^B c_j^{(b)}$, $j = 1, \dots, q$. Set $\hat{h} = \bar{c} S_{U^c} n^{-1/5}$ and $\hat{\lambda}_j = \bar{c}_j S_{U^c} n^{-2/5}$, where S_{U^c} is the standard deviation of U_i^c based on all n observations.

We will use \hat{h} and $\hat{\lambda}_j$, $j = 1, \dots, q$, in our empirical applications, where the single continuous variable U^c is *Experience* and U^d is composed of 6 categorical variables. We choose $m = 400$ and $B = 200$ below. When there are more than one continuous regressor in U , one can modify the above procedure correspondingly.

3 Monte Carlo Simulations

We now conduct Monte Carlo experiment to illustrate the finite sample performance of our nonparametric functional coefficient estimators with mixed data. In addition to the proposed estimator, we

also include several other parametric and nonparametric estimators.

The first data generating process (DGP) we consider is given by

$$Y_i = 0.1 (U_{i1}^2 + U_{i2} + U_{i3}) + 0.1 (U_{i1}U_{i2} + U_{i3}) X_{i1} + 0.15 (U_{i1}U_{i2} + U_{i3}) X_{i2} + \varepsilon_i,$$

where X_{ij} - Uniform(0, 4) ($j = 1, 2$), U_{i1} - Uniform(0, 4), $U_{ij} \in \{0, 1, \dots, 5\}$ with $P(U_{ij} = l) = 1/6$ for $l = 0, 1, \dots, 5$ and $j = 2, 3$, and ε_i - $N(0, 1)$. Furthermore, X_{ij} , U_{ij} , and ε_i are i.i.d. and mutually independent.

We consider two nonparametric estimators and three parametric estimators for the conditional mean function $m(x, u) = E(Y_i | X_i = x, U_i = u)$. We first obtain our nonparametric functional coefficient estimator (NP) with mixed data where the smoothing parameters (h, λ) are chosen by the least squares cross-validation. Then we obtain the nonparametric frequency estimator (NP-FREQ) with mixed data by using the cross-validated h and setting $\lambda = 0$ (see Li and Racine, 2007, Chapter 3). It is expected that the smaller the ratio of the sample size to the number of "cells", the worse the nonparametric frequency approach relative to our proposed kernel approach.

For the parametric estimation, we consider in practice what an applied econometrician would do when he or she confronts the data $\{(Y_i, X_i, U_i), 1 \leq i \leq n\}$ and have a strong belief that all the variables in X_i and U_i can affect the dependent variable Y_i . In the first parametric model, we ignore the potential interaction between regressors and estimate a linear model without any interaction (LIN) by regressing Y_i on X_i , U_{i1} , and the two categorical variables U_{i2} and U_{i3} . In the second parametric model, we take into account potential interaction between X_i and U_{1i} , and estimate a linear model with interaction (LIN-INT1) by adding the interaction terms between X_i and U_{1i} into the LIN model. In the third parametric model, we also consider the interaction between X_i and (U_{2i}, U_{3i}) , so we estimate a linear model with interaction (LIN-INT2) by adding the interaction terms between X_i and (U_{1i}, U_{2i}, U_{3i}) into the LIN-INT2 model. We expect LIN-INT2 outperforms LIN-INT1, which in turn outperforms LIN in terms of mean squared errors.

For performance measure, we will generate $2n$ observations $\{(Y_i, X_i, U_i), 1 \leq i \leq 2n\}$ for $n = 100, 200$, and 400 , and use the first n observations for in-sample estimation and evaluation, and the last n observations for out-of-sample evaluation. We consider root-mean-square error (RMSE) for both in-sample and out-of-sample evaluation:

$$RMSE_{in} = \sqrt{\frac{1}{n} \sum_{i=1}^n \{m(X_i, U_i) - \hat{m}(X_i, U_i)\}^2},$$

$$RMSE_{out} = \sqrt{\frac{1}{n} \sum_{i=1}^n \{m(X_{n+i}, U_{n+i}) - \hat{m}(X_{n+i}, U_{n+i})\}^2 M(X_{n+i}, U_{n+i}^c)},$$

where, for each method introduced earlier, $\hat{m}(x, u)$ is an estimate of $m(x, u)$ using the first n observations $\{(Y_i, X_i, U_i), 1 \leq i \leq n\}$, and $M(\cdot, \cdot)$ is a weight function for the out-of-sample evaluation. We use the weight function here because the out-of-sample observations can lie outside the data range of the in-sample observations, and when this occurs, the nonparametric methods significantly deteriorate. In this simulation study, we set $M(X_{n+i}, U_{c,n+i}) = \prod_{j=1}^{d+p} \mathbf{1}(|V_{ij} - \bar{V}_j| \leq 1.5s_{V_j})$, where $V_i = (X'_{n+i},$

Table 1: Comparison of finite sample performance of various estimators (DGP1)

n	Model	In-sample RMSE				Out-of-sample RMSE			
		Mean	Median	Std dev	IQR	Mean	Median	Std dev	IQR
100	NP	0.729	0.703	0.160	0.194	1.140	1.096	0.449	0.271
	NP-FREQ	0.993	0.994	0.141	0.193	5.494	3.222	25.067	1.374
	LIN	2.375	2.336	0.521	0.713	2.752	2.684	0.693	0.754
	LIN-INT1	1.686	1.649	0.353	0.495	2.088	2.027	0.464	0.591
	LIN-INT2	1.091	1.072	0.209	0.273	1.531	1.494	0.351	0.427
200	NP	0.523	0.512	0.080	0.097	0.798	0.789	0.114	0.131
	NP-FREQ	0.880	0.876	0.101	0.134	14.325	4.638	67.586	5.410
	LIN	2.436	2.431	0.370	0.503	2.637	2.586	0.390	0.515
	LIN-INT1	1.726	1.726	0.244	0.346	1.941	1.912	0.282	0.363
	LIN-INT2	1.116	1.115	0.154	0.214	1.320	1.304	0.196	0.252
400	NP	0.385	0.374	0.076	0.057	0.591	0.582	0.070	0.078
	NP-FREQ	0.573	0.564	0.074	0.087	7.681	2.144	44.756	2.221
	LIN	2.472	2.460	0.2806	0.361	2.563	2.550	0.280	0.369
	LIN-INT1	1.760	1.757	0.180	0.240	1.860	1.857	0.194	0.259
	LIN-INT2	1.138	1.132	0.109	0.149	1.243	1.235	0.128	0.171

U_{n+i}^c), and \bar{V}_j and sv_j denote the sample mean and standard deviation of $\{V_{ij}, 1 \leq i \leq n\}$, respectively. We report the mean, median, standard error, and interquartile range of $RMSE$ over 1000 Monte Carlo replications.

Table 1 reports the results for all five regression models. We summarize some interesting findings in Table 1. First, our proposed nonparametric functional coefficient estimator dominates all the other parametric or nonparametric estimators in terms of both in-sample RMSE and out-of-sample RMSE. Second, in comparison with the parametric estimators the nonparametric frequency estimator behaves reasonably well in terms of in-sample RMSE but not out-of-sample RMSE. The out-of-sample performance of the nonparametric frequency estimator is not acceptable even when the sample size is 400, in which case the average number of observations per cell is about 11. Third, as the sample size increases, the in-sample RMSEs of both our nonparametric estimator and the nonparametric frequency estimator decrease, but at rate slower than the parametric $n^{-1/2}$ -rate as expected. The same is true for the out-of-sample RMSE of our nonparametric estimator. Fourth, the performance of the parametric estimators based on misspecified models may not improve as the sample size increases.

We now consider a second DGP which allows for weak data dependence between observations. The data are generated from the following DGP

$$Y_i = U_{i1} (U_{i1} + U_{i2} + U_{i3}) + U_{i1} (U_{i1} + U_{i2} + U_{i3}) X_i + \varepsilon_i,$$

where

$$\begin{aligned} X_i &= 0.5X_{i-1} + e_{i1}, \\ U_{i1} &= 0.5 + 0.5U_{i-1,1} + e_{i2}, \end{aligned}$$

Table 2: Comparison of finite sample performance of various estimators (DGP2)

n	Model	In-sample RMSE				Out-of-sample RMSE			
		Mean	Median	Std dev	IQR	Mean	Median	Std dev	IQR
100	NP	0.388	0.369	0.122	0.155	0.606	0.575	0.184	0.222
	NP-FREQ	0.491	0.453	0.169	0.224	5.448	0.944	57.263	1.306
	LIN	2.565	2.437	0.856	1.013	3.050	2.871	1.008	1.216
	LIN-INT1	1.999	1.898	0.633	0.789	2.476	2.373	0.7788	1.009
	LIN-INT2	0.400	0.384	0.140	0.171	0.551	0.505	0.249	0.268
200	NP	0.231	0.218	0.070	0.090	0.392	0.372	0.097	0.114
	NP-FREQ	0.266	0.243	0.096	0.122	1.623	0.408	17.987	0.246
	LIN	2.646	2.576	0.639	0.834	2.901	2.847	0.669	0.854
	LIN-INT1	2.059	2.009	0.469	0.550	2.332	2.273	0.534	0.653
	LIN-INT2	0.384	0.369	0.105	0.142	0.468	0.437	0.142	0.168
400	NP	0.129	0.122	0.039	0.046	0.256	0.247	0.049	0.055
	NP-FREQ	0.138	0.126	0.046	0.051	0.274	0.255	0.140	0.062
	LIN	2.690	2.624	0.464	0.630	2.826	2.778	0.439	0.566
	LIN-INT1	2.096	2.066	0.328	0.442	2.240	2.217	0.342	0.459
	LIN-INT2	0.380	0.371	0.076	0.100	0.418	0.411	0.085	0.108

$\varepsilon_i \sim N(0, 1)$, $e_{i1} \sim N(0, 1)$, and $e_{i2} \sim \text{Uniform}(-0.5, 0.5)$, $U_{ij} \in \{-1, 0, 1\}$ with $P(U_{ij} = l) = 1/3$ for $l = -1, 0, 1$ and $j = 2, 3$. Furthermore, e_{ij} ($j = 1, 2$), U_{i2} , U_{i3} , and ε_i are i.i.d. and mutually independent.

Like the case for DGP 1, we also consider two nonparametric estimators and three parametric estimators for the conditional mean function $m(x, u) = E(Y_i | X_i = x, U_i = u)$. We denote the corresponding regression models as NP, NP-FREQ, LIN, LIN-INT1, and LIN-INT2, respectively. We again consider the performance measure in terms of *RMSE* for both in-sample and out-of-sample evaluation and for $n = 100, 200$, and 400 . The only difference is that when we generate $\{X_i, U_{i1}\}$, we throw away the first 200 observations to avoid the starting-up effects. We report the mean, median, standard error, and interquartile range of *RMSE* over 1000 Monte Carlo replications in Table 2. The findings in Table 2 are similar to those in Table 1. One noticeable difference is that the out-of-sample performance of the nonparametric frequency estimator is not bad when $n = 400$ for this DGP. We conjecture this is due to the fact the average number of observations per cell ($400/9 \approx 44$) is not small in this case.

4 An Empirical Application: Estimating the Wage Equation

In this section, we apply our functional coefficient model to estimate a wage equation embedded in the framework of Mincer's (1974) human capital earning function. The basic Mincer wage function takes the form:

$$\log Y = \beta_0 + \beta_1 S + \beta_2 A + \beta_3 A^2 + \epsilon, \quad (4.1)$$

where Y is some measure of individual earnings, S is years of schooling and A is age or work experience. In spite of its simplicity, Mincer Equation captures the reality remarkably well (Card, 1999), and has been firmly established as a benchmark in labor economics. Concerning its specification, several extensions have been made to allow more general parametric functional forms (see Murphy and Welch, 1990). Further, a nonparametric analysis has been done in Ullah (1985) and Zheng (2000). And in practice, other control variables, such as indicators of gender, race, occupation, or marital status are routinely included in the wage equation when they are available. Nevertheless, the additive separability assumption of the standard Mincer equation may be too stringent. For instance, it ignores the possibility that higher education results in more return to seniority.^[1] Also, it is often of keen economic and policy interest to investigate the differentials among different gender and race groups, where the return to education or experience may differ substantially. Therefore, we intend to estimate the functional coefficient model of the following form:

$$\log Y = a_1(U) + a_2(U)S + \epsilon, \quad (4.2)$$

where Y and S are as defined above, and U is a vector of mixed variables including one continuous variable – age or work experience, and six categorical variables for gender, race, marital status, veteran status, industry, and geographic location. The specification of (4.2) enables us to both study the direct effects of variables in U flexibly, and investigate whether and how they influence the return to education. Some past literature has already suggested nonlinear relationship between seniority and wage beyond a quadratic form (Murphy and Welch, 1990, Ullah, 1985, Zheng, 2000), as well as the fact that rising return to education from the 1980s is more drastic in the younger cohorts than in the older ones (Card and Lemieux, 2001).

Our model is also suitable for analyzing the gender and racial wage differentials. In the study of discrimination, it is common practice to estimate a “gender/racial wage gap” or estimate wage equation in separate samples. (For a survey of race and gender in the labor market, see Altonji and Blank, 1999.) Here the limitation of application of the traditional nonparametric method is the fact that indicators for gender and race are discrete, a problem overcome in our model. Also, compared with estimating wage separately among gender-racial groups or the frequency approach, our approach utilizes the entire dataset, thus achieving efficiency gain. We can also explicitly address other supposedly complicated interaction effects between the variables of interest. Further, unlike a complete nonparametric specification, model (4.2) has the further advantage that it can be readily extended to instrument variable estimation (Cai et al., 2006), provided we have some reasonable instruments to correct the endogeneity in education. To keep our discussion focused, however, this aspect is not further explored in this paper.

The data utilized are drawn from March CPS data of the year 1990, 1995, 2000 and 2005. The earning variable is the weekly earning calculated from annual salary income divided by weeks of work, and deflated by the CPI (1982-1984=100). As usual, we exclude observations that are part-time workers, self-employed, over 65, under 18, or earn less than 50 dollars per week. All observations fall into 3 racial categories – white, Hispanic and otherwise, 4 geographic location categories – Northeast, Midwest, South and West, and 10 industrial categories. There are also three dichotomous variables

Table 3: Linear Wage Equation

Year	1990		1995		2000		2005	
	(1)	(2)	(2)	(2)	(3)	(3)	(4)	(4)
Education	0.098 ^a	(0.002)	0.107 ^a	(0.002)	0.105 ^a	(0.003)	0.107 ^a	(0.002)
Experience	0.029 ^a	(0.001)	0.036 ^a	(0.002)	0.029 ^a	(0.002)	0.031 ^a	(0.001)
Experience ²	-0.000 ^a	(0.000)	-0.001 ^a	(0.000)	-0.001 ^a	(0.000)	-0.001 ^a	(0.000)
Female	-0.309 ^a	(0.010)	-0.290 ^a	(0.010)	-0.279 ^a	(0.011)	-0.277 ^a	(0.008)
White	0.100 ^a	(0.013)	0.130 ^a	(0.013)	0.097 ^a	(0.013)	-0.098 ^a	(0.010)
Hispanic	0.034 ^c	(0.017)	0.040 ^c	(0.022)	0.033 ^c	(0.019)	0.034 ^b	(0.014)
Single	-0.087 ^a	(0.009)	-0.071 ^a	(0.010)	-0.097 ^a	(0.010)	-0.102 ^a	(0.008)
Veteran	-0.013	(0.013)	-0.049 ^a	(0.015)	-0.008	(0.016)	-0.031 ^b	(0.014)
Observations	12328		10834		10433		17466	
R^2	0.37		0.36		0.33		0.34	

Note: 1) Heteroskedasticity-robust standard errors in parentheses.

2) a , b and c stand for 1%, 5% and 10% significant levels, respectively.

3) 3 region indicators, 9 industry indicators and a constant in all specifications.

“Female”, “Veteran” and “Single”. Years of schooling are estimated by records of the highest educational degree attained and experience is approximated by *Age-Schooling-6*. Figure 1 plots wage against experience and years of schooling for the four years under our investigation. The left panel in Figure 1 suggests the linear relationship (if any) between experience and wage is weak whereas the right panel in Figure 1 suggests there is a positive relationship between years of schooling and wage.

As a comparison, we also estimate a simple linear wage function, a linear wage function with interacting covariates, and a partially linear model. The results are reported in Table 3, Table 4, and Table 5 (see also Figure 2), respectively.

Results in Table 3 are in conformity with some stylized effects in labor economics, including stable return to schooling in the 1990s (Card and DiNardo, 2002; Beaudry and Green, 2004), concavity in return to experience, falling gender-wage gaps (Altonji and Blank, 1999), etc. The returns to schooling appear to range from 9.8% to 10.7% for the data under our investigation. Nevertheless, the inadequacy of a simple linear separable model is made clear in Table 4, since most of the interaction items of the covariates are significantly different from zero. And many of them are of important economic implications, such as the higher return to education for female and higher return to experience for the white. And the goodness-of-fit of the model after accounting for the interaction effects has also increased modestly. Table 4 indicates the omission of these interaction terms may cause significant bias in the estimate of returns to schooling, and the bias can be as large as about 41% for year 2005 if we believe the linear model with interaction terms is correctly specified.

Another extension of equation (4.1) is to consider the partially linear model: $\log Y = m(\textit{Schooling}, \textit{Experience}) + Z'\beta + \epsilon$, where Z is a set of dummy variables, and education and experience enter the model nonparametrically. We use the local linear method to estimate this model which is in the spirit of Robinson (1988). A second order Epanechnikov kernel $w(v) = 0.75(1 - 0.2v^2)1(|v| \leq \sqrt{5})$ is used; and the bandwidth is chosen by a LSCV method. Given the large number of observations

Table 4: Linear Wage Equation with Interacted Regressors

Year	1990		1995		2000		2005	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Education	0.133 ^a	(0.007)	0.146 ^a	(0.007)	0.134 ^a	(0.008)	0.151 ^a	(0.006)
Experience	0.059 ^a	(0.003)	0.071 ^a	(0.003)	0.049 ^a	(0.004)	0.053 ^a	(0.003)
Experience ²	-0.001 ^a	(0.000)	-0.001 ^a	(0.000)	-0.001 ^a	(0.000)	-0.001 ^a	(0.000)
Female	-0.349 ^a	(0.061)	-0.379 ^a	(0.069)	-0.526 ^a	(0.074)	-0.353 ^a	(0.059)
White	0.039	(0.089)	0.091	(0.091)	-0.077	(0.106)	0.025	(0.082)
Hispanic	0.496 ^a	(0.098)	0.551 ^a	(0.114)	0.455 ^a	(0.111)	0.607 ^a	(0.086)
Single	-0.132 ^a	(0.013)	-0.128 ^a	(0.014)	-0.137 ^a	(0.014)	-0.155 ^a	(0.012)
Veteran	-0.024 ^c	(0.014)	-0.056 ^a	(0.015)	-0.010 ^a	(0.017)	-0.027 ^c	(0.015)
Education×Experience	-0.002 ^a	(0.000)	-0.002 ^a	(0.000)	-0.001 ^a	(0.000)	-0.001 ^a	(0.000)
Education×Female	0.014 ^a	(0.004)	0.016 ^a	(0.004)	0.022 ^a	(0.005)	0.009 ^b	(0.004)
Education×White	0.009	(0.006)	0.007	(0.006)	0.010	(0.007)	0.006	(0.006)
Education×Hispanic	-0.034 ^a	(0.007)	-0.035 ^a	(0.008)	-0.039 ^a	(0.008)	-0.046 ^a	(0.006)
White×Female	-0.135 ^a	(0.025)	-0.123 ^a	(0.026)	-0.087 ^a	(0.026)	-0.098 ^a	(0.020)
Hispanic×Female	-0.017	(0.034)	-0.069	(0.043)	-0.035	(0.038)	0.012	(0.028)
Single×Female	0.114 ^a	(0.018)	0.141 ^a	(0.018)	0.105 ^a	(0.020)	0.135 ^a	(0.010)
Experience×Female	-0.005 ^a	(0.001)	-0.004 ^a	(0.001)	-0.002 ^a	(0.001)	-0.001 ^a	(0.001)
Experience×White	0.000	(0.001)	0.000	(0.001)	0.004 ^a	(0.001)	0.002 ^a	(0.001)
Experience×Hispanic	-0.003 ^c	(0.002)	-0.004 ^b	(0.002)	0.001	(0.002)	-0.000	(0.001)
Observations	12328		10834		10433		17466	
R^2	0.39		0.38		0.34		0.36	

Note: 1) Heteroskedasticity-robust standard errors in parentheses.

2) a, b and c stand for 1%, 5% and 10% significant levels, respectively.

3) 3 region indicators, 9 industry indicators and a constant in all specifications.

Table 5: Partially Linear Wage Equation

Year	1990		1995		2000		2005	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female	-0.280 ^a	(0.010)	-0.265 ^a	(0.011)	-0.259 ^a	(0.011)	-0.259 ^a	(0.008)
White	0.103 ^a	(0.012)	0.135 ^a	(0.013)	0.096 ^a	(0.013)	0.102 ^a	(0.010)
Hispanic	0.001	(0.017)	-0.001 ^a	(0.022)	-0.017	(0.019)	-0.007	(0.014)
Single	-0.077 ^a	(0.009)	-0.058 ^a	(0.010)	-0.082 ^a	(0.010)	-0.077 ^a	(0.008)
Veteran	0.024 ^a	(0.013)	-0.009	(0.015)	0.021	(0.016)	-0.001	(0.014)
Observations	12328		10834		10433		17446	
R^2	0.40		0.39		0.36		0.38	

Note: 1) Heteroskedasticity-robust standard errors in parentheses.

2) a, b and c stand for 1%, 5% and 10% significant levels, respectively.

3) 3 region indicators, 9 industry indicators and a constant in all specifications.

4) The estimate of $m(\text{Schooling}, \text{Experience})$ is plotted in Figure 2.

in our dataset, it is extremely costly to apply the LSCV method directly on all the observations. So we apply a methodology similar to that proposed at the end of section 2 to choose the bandwidth. As reported in Table 5, the partially linear model performs a little bit better in goodness-of-fit, as expected. However, it is noteworthy that comparing with the simple linear model, accounting for the possibly complex function form of education and experience has also significantly changed the estimates of the coefficients for the other covariates. For instance, the effects of race have drastically dropped in magnitude as well as significance. The difference may be the result of biases induced by the misspecification in a parametric model, and thus indicates the needs for the more general functional form assumption.

In all the above specifications, we use dummy variables to allow different intercepts for different regions and industries, and the majority of them have a significant estimated coefficient. The large number of categories makes it difficult to study their interaction effects with other regressors. In contrast, in the nonparametric framework of mixed regressors, only one categorical variable is necessary to describe such characteristic as industry or location. And this advantage has made our proposed model further suitable for the application.

For a comprehensive presentation of the regression results of model (4.2), we plot the wage-experience profiles of different cells defined by a discrete characteristic averaged over other categorical covariates. We use the second order Epanechnikov kernel in our nonparametric estimation, and choose the bandwidth by the LSCV method introduced at the end of Section 2.4. The R^2 's of the model have been increased up to 0.66, 0.65, 0.62, 0.68, respectively for the four years.

Figure 3 reports the estimated $a_1(\textit{Experience}, \textit{Region}, :)$ and $a_2(\textit{Experience}, \textit{Region}, :)$ of model (4.2) for different regions averaged across all other categorical variables. $a_1(\textit{Experience}, \textit{Region}, :)$ can be viewed as the direct effects of experience on wage for the particular region (averaged across all other categorical variables), and $a_2(\textit{Experience}, \textit{Region}, :)$ represents the return to schooling as a function of experience for the particular region. We summarize some interesting findings from Figure 3. First, while there are considerable variations between regions, we find the direct effects of experience on wage are usually **positive** (upward sloping) but **not necessarily concave**, which is in sharp contrast with the results of the parametric model. Notably, the experience-wage profile estimated here are from cross-sections and cannot be taken as individuals life-cycle earning trend. Second, if the standard Mincer equation holds, we expect the estimated $a_2(\textit{Experience}, \textit{Region}, :)$ to be a horizontal line. But clearly, this is far from reality. The effects of experience on return to schooling are **mainly negative**, which agrees with our previous results from the parametric setting, presented in Table 4. The findings here have interesting econometric interpretation. On the one hand, we may wonder if higher education causes higher return to seniority, or similarly, longer experience leads to higher return to education. On the other hand, it is possible that the young cohorts (implied by shorter experience) have higher return to education, due to cohort supply effects, technological changes or some other reasons. And we need to resort to empirical results to evaluate the overall influence. In the sample studied here, the later force has been found to dominate the former in their direction of impacts. Admittedly, the interacting patterns of the regressors in the wage equation uncovered by this functional coefficient model require further careful investigation.

Figure 4 reports the estimated $a_1(\textit{Experience}, \textit{Race}, :)$ and $a_2(\textit{Experience}, \textit{Race}, :)$ of model (4.2) for different races averaged across all other categorical variables. $a_1(\textit{Experience}, \textit{Race}, :)$ can be viewed as the direct effects of experience on wage for the race, and $a_2(\textit{Experience}, \textit{Race}, :)$ represents the return to schooling as a function of experience for the particular race. The findings are similar to those in Figure 3. We only mention that the return to schooling seems much higher for White and others (above 0.1 across 2/3 of the range of experience) than Hispanic (below 0.1 in almost all the range of experience).

Figure 5 reports the estimated $a_1(\textit{Experience}, :)$ and $a_2(\textit{Experience}, :)$ depending on whether a person is male or female, single or non-single, and veteran or non-veteran. Figure 6 reports the estimated $a_1(\textit{Experience}, \textit{Industry}, :)$ and $a_2(\textit{Experience}, \textit{Industry}, :)$ of model (4.2) for different industries averaged across all other categorical variables. Both figures can be interpreted similarly to the case of Figure 3. The most eminent implication by these figures is that return to education does depend heavily upon other variables. In particular, the top panel in Figure 5 indicates that higher return to education for female across all the range of age or work experience. In addition, we can see substantial variation among the cells which suggests the highly complex functional form of the wage equation.

Figure 7 reports the estimated $a_1(\textit{Experience}, :)$ and $a_2(\textit{Experience}, :)$ averaged over all categorical variables. Similarly to the cases of Figures 3-6, we observe that the direct impact of experience on wage is **positive** but the return to schooling as a function of experience tends to be decreasing except when experience is low (≤ 4 years in 1990, ≤ 12 in 2005). When experience is larger than 37 years, the return to schooling is diminishing very fast a function of experience. Prior to 37 years, the returns to schooling may vary from 0.105 to 0.145.

Therefore, our empirical application has demonstrated the usefulness of our proposed model in uncovering complicated patterns of interacting effects of the covariates on the dependent variable. And the results are of interesting economic interpretation.

5 Conclusions

This paper proposes a local linear functional coefficient estimator that admits a mix of discrete and continuous data for stationary time series. Under weak conditions our estimator is asymptotically normally distributed. We also include simulations and empirical applications. We find from the simulations that our nonparametric estimators behave reasonably well for a variety of DGPs.

As an empirical application, we estimate a human capital earning function from the recent CPS data. Unlike the widely used linear separable model, or the frequency approach that conducts estimation in splitted samples, the proposed model enables us to utilize the entire dataset and allows the return to education to vary with the other categorical and continuous variables. The empirical findings show considerable interacting effects among the regressors in the wage equation. For instance, the younger cohorts are found to have higher return to education. While these patterns need further explanation from labor economic theory, the application demonstrates the usefulness of our proposed functional coefficient model due to its flexibility and clear economic interpretation. And

thus the model has good potential for applied research. Our future research will address some related problems such as the optimal selection of smoothing parameters. Another extensions is to study the estimation of functional coefficient model with both endogeneity and mixed regressors.

Acknowledgements

The authors gratefully thank the editors and two anonymous referees for their constructive comments and suggestions. They also thank Zongwu Cai for his helpful comment on an early version of this paper. Correspondence: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore, 178903 (E-mail: ljsu@smu.edu.sg), Phone: +65 6828-0386. Ye Chen, Department of Economics, Princeton University, Princeton, NJ 08544-1021 (E-mail: yechen@Princeton.EDU). Aman Ullah, Department of Economics, UCR, Riverside, CA 92521-0427 (E-mail: aman.ullah@ucr.edu; Phone: 951-827-1591). The first author gratefully acknowledges financial support from the NSFC (Project 70501001 and 70601001). The third author gratefully acknowledges the financial support from the Academic Senate, UCR.

6 Appendix

We use $\|\cdot\|$ to denote the Euclidean norm of \cdot , C to signify a generic constant whose exact value may vary from case to case, and a' to denote the transpose of a . Let $d_{u_i u} = \sum_{t=1}^q \mathbf{1}(U_{it}^d \neq u_t^d)$, where $\mathbf{1}(U_{it}^d \neq u_t^d)$ is an indicator function that takes value 1 if $U_{it}^d \neq u_t^d$ and 0 otherwise. So $d_{u_i u}$ indicates the number of disagreeing components between U_{it}^d and u_t^d .

Proof of Theorem 2.1

We first define some notation. For any $p \times 1$ vectors $c = (c_1, \dots, c_p)'$ and $d = (d_1, \dots, d_p)'$, let $c \odot d = (c_1 d_1, \dots, c_p d_p)'$ and $c/d = (c_1/d_1, \dots, c_p/d_p)'$ whenever applicable. Let

$$S_n = S_n(u) = \begin{pmatrix} S_{n,0} & S_{n,1} \\ S'_{n,1} & S_{n,2} \end{pmatrix}, \quad T_n = T_n(u) = T_{n,1} + T_{n,2}, \quad (6.1)$$

with

$$\begin{aligned}
S_{n,0} &= S_{n,0}(u) = n^{-1} \sum_{i=1}^n X_i X_i' K_{iu}, \\
S_{n,1} &= S_{n,1}(u) = n^{-1} \sum_{i=1}^n \left(X_i X_i' \right) \otimes \left((U_i^c - u^c) / h \right)' K_{iu}, \\
S_{n,2} &= S_{n,2}(u) = n^{-1} \sum_{i=1}^n \left(X_i X_i' \right) \otimes \left(((U_i^c - u^c) / h) \left((U_i^c - u^c) / h \right)' \right) K_{iu}, \\
T_{n,1} &= T_{n,1}(u) = n^{-1} \sum_{i=1}^n \left(\begin{array}{c} X_i \varepsilon_i \\ (X_i \varepsilon_i) \otimes \left((U_i^c - u^c) / h \right) \end{array} \right) K_{iu}, \text{ and} \\
T_{n,2} &= T_{n,2}(u) = n^{-1} \sum_{i=1}^n \left(\begin{array}{c} (X_i X_i' \mathbf{a}(U_i)) \\ (X_i X_i' \mathbf{a}(U_i)) \otimes \left((U_i^c - u^c) / h \right) \end{array} \right) K_{iu},
\end{aligned}$$

where recall $\mathbf{a}(U_i) = (a_1(U_i), \dots, a_d(U_i))'$. Then

$$\hat{\theta} = H_1^{-1} S_n^{-1} T_n,$$

where $H_1 = \text{diag}(1, \dots, 1, h', \dots, h')$ is a $d(p+1) \times d(p+1)$ diagonal matrix with d diagonal elements of 1 and d diagonal elements of h . Let $H = \sqrt{nh_1 \dots h_p}$. Then

$$\begin{aligned}
HH_1 (\hat{\theta} - \theta) &= HS_n^{-1} (T_n - S_n \theta) \\
&= HS_n^{-1} T_{n,1} + HS_n^{-1} (T_{n,2} - S_n \theta).
\end{aligned}$$

We first prove several lemmas.

Lemma 6.1 (a) $S_{n,0} = \Omega(u) f_u(u) + o_p(1)$,
(b) $S_{n,1} = O_p(\|h\|^2 + \|h\| \|\lambda\|) = o_p(1)$,
(c) $S_{n,2} = \mu_{2,1}(\Omega(u) f_u(u)) \otimes I_p + o_p(1)$.

Proof. We only prove (a) since the proofs of (b) and (c) are similar. First by the stationarity of $\{X_i, U_i\}$,

$$\begin{aligned}
E(S_{n,0}) &= E\left(X_i X_i' K_{iu}\right) \\
&= E\left(X_i X_i' W_{h,iu} | d_{u_i u} = 0\right) p(u^d) + \sum_{s=1}^q E\left(X_i X_i' W_{h,iu} L_{\lambda,iu} | d_{u_i u} = s\right) P(d_{u_i u} = s) \\
&= E(\Omega(U_i) W_{h,iu} | d_{u_i u} = 0) p(u^d) + O(\|\lambda\|) \\
&= \int \Omega(u^c + h \odot v, u^d) f_u(u^c + h \odot v, u^d) W(v) dv + O(\|\lambda\|) \\
&= \Omega(u) f_u(u) + O(\|h\|^2 + \|\lambda\|), \tag{6.2}
\end{aligned}$$

where $p(u^d) = P(U_i^d = u^d)$.

Since a typical element of $S_{n,0}$ is

$$s_{n,st} = n^{-1} \sum_{i=1}^n X_{is} X_{it} K_{iu}, \quad s, t = 1, \dots, d,$$

by the Chebyshev's inequality, it suffices to show that

$$\text{var}(s_{n,st}) = o(1). \quad (6.3)$$

Let $\xi_i = X_{is} X_{it} K_{iu}$. By the stationarity of $\{X_i, U_i\}$, we have

$$\text{var}(s_{n,st}) = \frac{1}{n} \text{var}(\xi_1) + \frac{2}{n} \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \text{cov}(\xi_1, \xi_j). \quad (6.4)$$

Clearly,

$$\text{var}(\xi_1) \leq E(X_{1s}^2 X_{1t}^2 K_{1u}^2) = O\left((h_1 \dots h_n)^{-1}\right). \quad (6.5)$$

To obtain an upper bound for the second term on the right hand side of (6.4), we split it into two terms as follows

$$\sum_{j=1}^{n-1} |\text{cov}(\xi_1, \xi_j)| = \sum_{j=1}^{d_n} |\text{cov}(\xi_1, \xi_j)| + \sum_{j=d_n+1}^{n-1} |\text{cov}(\xi_1, \xi_j)| \equiv J_1 + J_2,$$

where d_n is a sequence of positive integers such that $d_n h_1 \dots h_p \rightarrow 0$ as $n \rightarrow \infty$. Since for any $j > 1$,

$$|E(\xi_1 \xi_j)| = |E(X_{1s} X_{1t} K_{1,u} X_{js} X_{jt} K_{j,u})| = O(1),$$

$J_1 = O(d_n)$. For J_2 , by the Davydov's inequality (e.g., Hall and Heyde, 1980, p. 278; or Bosq, 1996, p. 19), we have

$$\begin{aligned} \text{cov}(\xi_1, \xi_j) &\leq C [\alpha(j-1)]^{\gamma/(2+\gamma)} \left(E|\xi_1|^{2+\gamma}\right)^{2/(2+\gamma)} \\ &= C [\alpha(j-1)]^{\gamma/(2+\gamma)} \left\{E\left|(X_{1s} X_{1t})^{(2+\gamma)} K_{1,u}^{2+\gamma}\right|\right\}^{2/(2+\gamma)} \\ &= O\left((h_1 \dots h_p)^{-(2+2\gamma)/(2+\gamma)}\right) [\alpha(j-1)]^{\gamma/(2+\gamma)}. \end{aligned} \quad (6.6)$$

So

$$\begin{aligned} J_2 &\leq C (h_1 \dots h_p)^{-(2+2\gamma)/(2+\gamma)} \sum_{j=d_n}^{n-1} [\alpha(j)]^{\gamma/(2+\gamma)} \\ &\leq C (h_1 \dots h_p)^{-(2+2\gamma)/(2+\gamma)} d_n^{-\alpha} \sum_{j=d_n}^{\infty} j^\alpha [\alpha(j)]^{\gamma/(2+\gamma)} = o\left((h_1 \dots h_p)^{-1}\right), \end{aligned} \quad (6.7)$$

by choosing d_n such that $d_n^{-\alpha} (h_1 \dots h_p)^{-\gamma/(2+\gamma)} = o(1)$. This, in conjunction with (6.4)-(6.5), implies, $\text{var}(s_{n,st}) = O\left((n h_1 \dots h_p)^{-1}\right) = o(1)$. ■

Lemma 6.2

$$HT_{n,1} = n^{-1/2} (h_1 \dots h_p)^{1/2} \sum_{i=1}^n \begin{pmatrix} X_i \varepsilon_i \\ (X_i \varepsilon_i) \otimes ((U_i^c - u^c)/h) \end{pmatrix} K_{iu} \xrightarrow{d} N(0, \Gamma),$$

where $H = \sqrt{nh_1 \dots h_p}$, $\sigma^2(u, x) = E[\varepsilon_i^2 | U_i = u, X_i = x]$, $\Omega^*(u) = E[X_i X_i' \sigma^2(U_i, X_i) | U_i = u]$, and

$$\Gamma = \Gamma(u) = f_u(u) \begin{pmatrix} \mu_{0,2}^p \Omega^*(u) & 0' \\ 0 & \mu_{2,2} \Omega^*(u) \otimes I_p \end{pmatrix}.$$

Proof. Let ω be a unit vector on $\mathbb{R}^{d(p+1)}$. Let

$$\zeta_i = (h_1 \dots h_p)^{1/2} \omega' \begin{pmatrix} X_i \varepsilon_i \\ (X_i \varepsilon_i) \otimes ((U_i^c - u^c)/h) \end{pmatrix} K_{iu}.$$

By the Cramér-Wold device, it suffices to prove

$$I_n = n^{-1/2} \sum_{i=1}^n \zeta_i \xrightarrow{d} N(0, \omega' \Gamma \omega). \quad (6.8)$$

Clearly, by the law of iterated expectation, $E(\zeta_i) = 0$. Now

$$\text{var}(I_n) = \text{var}(\zeta_1) + 2 \sum_{j=1}^{n-1} \left(1 - \frac{j}{n}\right) \text{cov}(\zeta_1, \zeta_j).$$

By arguments similar to those used in the proof of Lemma 6.1,

$$\begin{aligned} & \text{var}(\zeta_1) \\ &= h_1 \dots h_p \omega' E \left\{ \begin{pmatrix} \Omega^*(U_i) & \Omega^*(U_i) \otimes ((U_i^c - u^c)/h)' \\ \Omega^*(U_i) \otimes ((U_i^c - u^c)/h) & \Omega^*(U_i) \otimes (((U_i^c - u^c)/h)((U_i^c - u^c)/h)') \end{pmatrix} K_{iu}^2 \right\} \omega \\ &= \omega' \Gamma \omega + o(1), \end{aligned}$$

and

$$\sum_{j=1}^{n-1} |\text{cov}(\zeta_1, \zeta_j)| = o(1),$$

which implies that

$$\text{var}(I_n) \rightarrow \omega' \Gamma \omega \text{ as } n \rightarrow \infty.$$

Using the standard Doob's small-block and large-block technique, we can finish the rest of the proof by following the arguments of Cai, Fan and Yao (2000, pp. 954-955) or Cai and Ould-Saïd (2003, pp. 446-448). ■

Lemma 6.3 Let $B_n = H(T_{n,2} - S_n \theta)$. Then $B_n = b(h, \lambda) + o_p(1)$, where $b(h, \lambda)$ is defined in (2.9).

Proof. Let

$$\begin{aligned} \varsigma_i &= H \begin{pmatrix} (X_i X_i' \mathbf{a}(U_i)) \\ (X_i X_i' \mathbf{a}(U_i)) \otimes ((U_i^c - u^c)/h) \end{pmatrix} K_{iu} \\ &\quad - H \begin{pmatrix} (X_i X_i') & (X_i X_i') \otimes ((U_i^c - u^c)/h)' \\ (X_i X_i') \otimes ((U_i^c - u^c)/h) & (X_i X_i') \otimes (((U_i^c - u^c)/h)((U_i^c - u^c)/h)') \end{pmatrix} \theta K_{iu}. \end{aligned}$$

Then we have

$$B_n = \frac{1}{n} \sum_{i=1}^n \varsigma_i. \quad (6.9)$$

Let $\bar{\varsigma}_i = E(\varsigma_i | U_i)$. Then

$$\begin{aligned} E(B_n) &= E(\bar{\varsigma}_i) \\ &= E\{\bar{\varsigma}_i | d_{u_i u} = 0\} p(u^d) + E\{\bar{\varsigma}_i | d_{u_i u} = 1\} P(d_{u_i u} = 1) + O(H \|h\|^2) \\ &\equiv b_{n,1} + b_{n,2} + o(1). \end{aligned}$$

On the set $\{U_i^d = u^d, W_{h,iu} > 0\}$,

$$a_j(U_i) = a_j(u) + \dot{a}_j(u)' (U_i^c - u^c) + \frac{1}{2} (U_i^c - u^c)' \ddot{a}_j(u) (U_i^c - u^c) + o(\|h\|^2).$$

Let $A(U_i, u) = ((U_i^c - u^c)' \ddot{a}_1(u) (U_i^c - u^c), \dots, (U_i^c - u^c)' \ddot{a}_d(u) (U_i^c - u^c))'$. Recall $A = (\sum_{s=1}^p h_s^2 a_{1,ss}(u), \dots, \sum_{s=1}^p h_s^2 a_{d,ss}(u))'$, and $\mathbf{b}(u) = (\dot{a}_1(u)', \dots, \dot{a}_d(u)')$. Then we have

$$\begin{aligned} b_{n,1} &= \frac{1}{2} H E \left\{ \begin{pmatrix} \Omega(U_i) A(U_i, u) \\ (\Omega(U_i) A(U_i, u)) \otimes ((U_i^c - u^c)/h) \end{pmatrix} W_{h,iu} \Big| d_{u_i u} = 0 \right\} \times p(u^d) + o(1) \\ &= \frac{H \mu_{2,1}}{2} \begin{pmatrix} f_u(u) \Omega(u) A \\ 0 \end{pmatrix} + o(1), \end{aligned}$$

and

$$\begin{aligned} &b_{n,2} \\ &= H E \{ \bar{\varsigma}_i | d_{u_i u} = 1 \} P(d_{u_i u} = 1) \\ &= H E \left\{ \begin{pmatrix} \Omega(U_i) (\mathbf{a}(U_i) - \mathbf{a}(u)) - (\Omega(U_i) \otimes ((U_i^c - u^c)/h)') \mathbf{b}(u) \\ (\Omega(U_i) (\mathbf{a}(U_i) - \mathbf{a}(u))) \otimes ((U_i^c - u^c)/h) - (\Omega(U_i) \otimes (((U_i^c - u^c)/h) ((U_i^c - u^c)/h)')) \mathbf{b}(u) \end{pmatrix} \right. \\ &\quad \left. \times K_{iu} \Big| d_{u_i u} = 1 \right\} P(d_{u_i u} = 1) + o(1) \\ &= H \sum_{\tilde{u}^d \in \mathcal{D}} \sum_{s=1}^q \lambda_s I_s(u^d, \tilde{u}^d) f_u(u^c, \tilde{u}^d) \begin{pmatrix} \Omega(u^c, \tilde{u}^d) (\mathbf{a}(u^c, \tilde{u}^d) - \mathbf{a}(u)) \\ -\mu_{2,1} (\Omega(u^c, \tilde{u}^d) \otimes I_p) \mathbf{b}(u) \end{pmatrix} + o(1). \end{aligned}$$

Consequently, $E(B_n) = b(h, \lambda) + o(1)$, where $b(h, \lambda)$ is defined in (2.9).

To show $\text{var}(B_n) = o(1)$ elementwise, we focus on the first d elements $\varsigma_i^{(1)}$ of ς_i since the other cases are similar, where

$$\varsigma_i^{(1)} = H [X_i X_i' (\mathbf{a}(U_i) - \mathbf{a}(u)) - (X_i X_i' \otimes ((U_i^c - u^c)/h)') \mathbf{b}(u)] K_{iu}.$$

A typical element of $\varsigma_i^{(1)}$ is

$$\varsigma_{i,t}^{(1)} = H \left[X_{it} \sum_{s=1}^d X_{is} (a_s(U_i) - a_s(u)) - X_{it} \sum_{s=1}^d X_{is} ((U_i^c - u^c)/h)' b_j(u) \right] K_{iu},$$

$t = 1, \dots, d$.

$$\text{var} \left(\frac{1}{n} \sum_{i=1}^n \varsigma_{i,t}^{(1)} \right) = \frac{1}{n} \text{var} \left(\varsigma_{1,t}^{(1)} \right) + \frac{2}{n} \sum_{j=1}^{n-1} \left(1 - \frac{j}{n} \right) \text{cov} \left(\varsigma_{1,t}^{(1)}, \varsigma_{j,t}^{(1)} \right).$$

By arguments similar to those used in the proof of Lemma 6.1,

$$\frac{1}{n} \text{var} \left(\varsigma_{1,t}^{(1)} \right) = O \left(\|h\|^4 + \|\lambda\|^2 \right) = o(1),$$

and

$$\sum_{j=1}^{n-1} \left| \text{cov} \left(\varsigma_{1,t}^{(1)}, \varsigma_{j,t}^{(1)} \right) \right| = o(1),$$

which implies that $\text{var} \left(\frac{1}{n} \sum_{i=1}^n \varsigma_{i,t}^{(1)} \right) = o(1)$. Similarly, one can show that the variance of the other elements in B_n is $o(1)$. The conclusion then follows by the Chebyshev's inequality. ■

By Lemmas 6.1-6.3,

$$HH_1 \left(\hat{\theta} - \theta \right) - B^{-1}b(h, \lambda) \xrightarrow{d} N \left(0, B^{-1}\Gamma B^{-1} \right).$$

This completes the proof of Theorem 2.1.

Proof of Corollary 2.2

Since the proof parallels that of Theorem 2.1, we only sketch the difference. Recall $S_n(u)$ is defined in (6.1). When $u^c = \nu h$, we have

$$\begin{aligned} E[S_{n,0}(u)] &= E \left[X_i X_i' K_{iu} \right] \\ &= \int \Omega(u^c + hz, u^d) f_u(u^c + hz, u^d) W(z) dz + O(\|\lambda\|) \\ &= \Omega(0, u^d) f_u(0, u^d) \iota_{\nu 0} + o(1), \end{aligned}$$

where $\iota_{\nu 0}$ is defined after (2.10). Similarly, $E[S_{n,1}(u)] = \Omega(0, u^d) f_u(0, u^d) \iota_{\nu 1} + o(1)$, and $E[S_{n,2}(u)] = \Omega(0, u^d) f_u(0, u^d) \iota_{\nu 2} + o(1)$. It follows that

$$S_n(u^c, u^d) \xrightarrow{p} S_\nu \otimes \Omega(0, u^d) f_u(0, u^d), \quad (6.10)$$

where S_ν is defined in (2.10). Following the proof of Lemma 6.2, with $u^c = \nu h$ we can show that

$$\text{var}(HT_{n,1}) = \Gamma_\nu \otimes \Omega^*(0, u^d) f_u(0, u^d) + o(1), \quad (6.11)$$

where Γ_ν is defined in (2.10). Following the proof of Lemma 6.3, when $u^c = \nu h$, we have

$$\begin{aligned} b_{n,1} &= \frac{1}{2} H E \left\{ \left(\begin{array}{c} \Omega(U_i) A(U_i, u) \\ \Omega(U_i) A(U_i, u) ((U_i^c - u^c)/h) \end{array} \right) W_{h,iu} \Big| d_{u,iu} = 0 \right\} \times p(u^d) + o(1) \\ &= \frac{H}{2} \left(\begin{array}{c} \Omega(0, u^d) \bar{A}(0, u^d) \iota_{\nu 2} \\ \Omega(0, u^d) \bar{A}(0, u^d) \iota_{\nu 3} \end{array} \right) f_u(0, u^d) + o(1), \end{aligned} \quad (6.12)$$

and

$$\begin{aligned}
b_{n,2} &= H E \left\{ \left(\begin{array}{l} \Omega(U_i) (\mathbf{a}(U_i) - \mathbf{a}(u)) - ((U_i^c - u^c)/h) \Omega(U_i) \mathbf{b}(u) \\ \Omega(U_i) (\mathbf{a}(U_i) - \mathbf{a}(u)) ((U_i^c - u^c)/h) - ((U_i^c - u^c)/h)^2 \Omega(U_i) \mathbf{b}(u) \end{array} \right) K_{iu} \Big|_{d_{u_i u}=1} \right\} \\
&\quad \times P(d_{u_i u} = 1) + o(1) \\
&= H \sum_{\tilde{u}^d \in \mathcal{D}} \sum_{s=1}^q \lambda_s I_s(u^d, \tilde{u}^d) f_u(0, \tilde{u}^d) \left(\begin{array}{l} \Omega(0, \tilde{u}^d) \{ \iota_{\nu 0} [\mathbf{a}(0, \tilde{u}^d) - \mathbf{a}(0, u^d)] - \iota_{\nu 1} \mathbf{b}(0, u^d) \} \\ \Omega(0, \tilde{u}^d) \{ \iota_{\nu 1} [\mathbf{a}(0, \tilde{u}^d) - \mathbf{a}(0, u^d)] - \iota_{\nu 2} \mathbf{b}(0, u^d) \} \end{array} \right) \\
&\quad + o(1), \tag{6.13}
\end{aligned}$$

where $\bar{A}(0, u^d)$ is defined in (2.11). Combining (6.10)-(6.13) yields the desired result. ■

Notes

[1] Throughout our paper the use of word return or marginal return from education refers to the functional (varying) coefficient of education which may not be the marginal return if the education is endogenous, an issue not explored in our paper.

References

- Aitchison, J., and Aitken, C. G. G. (1976), “Multivariate binary discrimination by the kernel method”, *Biometrika* 63, pp. 413-420.
- Altonji, J. G., and Blank, R.M. (1999), “Race and gender in the labor market”, in O. C. Ashenfelter and D. Card (Eds), *Handbook of Labor Economics 3C*, Elsevier, North Holland, Chapter 48, pp. 3143-3259.
- Beaudry, P., and Green, D. A. (2004), “Changes in US wages, 1976-2000: ongoing skill bias or major technological change?” *Journal of Labor Economics* 23, pp. 491-526.
- Bosq, D. (1996), *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction*. Springer, New York.
- Cai, Z., M. Das, Xiong, H., and Wu, X. (2006), “Functional coefficient instrumental variables models”, *Journal of Econometrics* 133, pp. 207-241.
- Cai, Z., J. Fan, and Yao, Q. (2000), “Functional-coefficient regression models for nonlinear time series”, *Journal of American Statistical Association* 95, pp. 941-956.
- Cai, Z., and Ould-Saïd, E. (2003), “Local M-estimator for nonparametric time series”, *Statistics and Probability Letters* 65, pp. 433-449.
- Card, D. (1999), “Casual effect of education on earnings”, in O. C. Ashenfelter and D. Card (Eds), *Handbook of Labor Economics 3A*, Elsevier, North Holland, Chapter 48, pp. 1802-1864.
- Card, D., and DiNardo, J. (2002), “Skill biased technological change and rising wage inequality: some problems and puzzles”, *Journal of Labor Economics* 20, pp. 733-783.

- Card, D., and Lemieux, T. (2001), "Can falling supply explain the rising return to college for younger men? A cohort-based analysis", *The Quarterly Journal of Economics* 116, pp. 705-746.
- Chen, R. and Tsay, R. S. (1993), "Functional-coefficient autoregressive models", *Journal of American Statistical Association* 88, pp. 298-308.
- Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992), "Local regression models", in J. M. Chambers and T. J. Hastie (Eds), *Statistical Models in S*, Pacific Grove, CA, Wadsworth & Brooks/Cole, pp. 309-376.
- Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, vol. 66 of Monographs on Statistics and Applied Probability. Chapman and Hall.
- Fan, Y. and Li, Q. (1999), "Root- n -consistent estimation of partially linear time series models", *Journal of Nonparametric Statistics* 11, pp. 251-269.
- Hall, P. and Heyde, C. C. (1980), *Martingale Limit Theory and Its Applications*. Academic Press, New York.
- Hall, P., Wolf, R. C. L., and Yao, Q. (1999), "Methods of estimating a conditional distribution function", *Journal of the American Statistical Association* 94, pp. 154-163.
- Hastie, T. J., and Tibshirani, R. J. (1993), "Varying-coefficient models (with discussion)", *Journal of the Royal Statistical Society, Series B.* 55, pp. 757-796.
- Li, Q., and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton and Oxford.
- Li, Q., and Racine, J. (2008a), "Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data", *Journal of Business and Economic Statistics* 26, pp. 423-434.
- Li, Q., and Racine, J. (2008b), "Smoothing Varying-Coefficient Estimation and Inference for Qualitative and Quantitative Data", Mimeo. Dept. of Economics, Texas A&M University.
- Mincer, J. (1974), *Schooling, Experience and Earnings*, New York, National Bureau of Economic Research.
- Murphy, K., and Welch, F. (1990), "Empirical age-earnings profiles", *Journal of Labor Economics* 8, pp. 202-229.
- Racine, J. and Li, Q. (2004), "Nonparametric estimation of regression functions with both categorical and continuous data", *Journal of Econometrics* 119, pp. 99-130.
- Robinson, P. M. (1988), "Root- n -consistent semiparametric regression", *Econometrica* 56, pp. 931-954.

Ullah, A. (1985), "Specification analysis of econometric models", *Journal of Quantitative Economics* 1, pp. 187-209.

Zheng, J. (2000), "Specification testing and nonparametric estimation of the human capital model", in T. B. Fomby & R. C. Hill (Eds), *Applying Kernel and Nonparametric Estimation to Economic Topics (Advances in Econometrics 14)*, JAI Press Inc, Stamford, Conn. pp. 129-154.

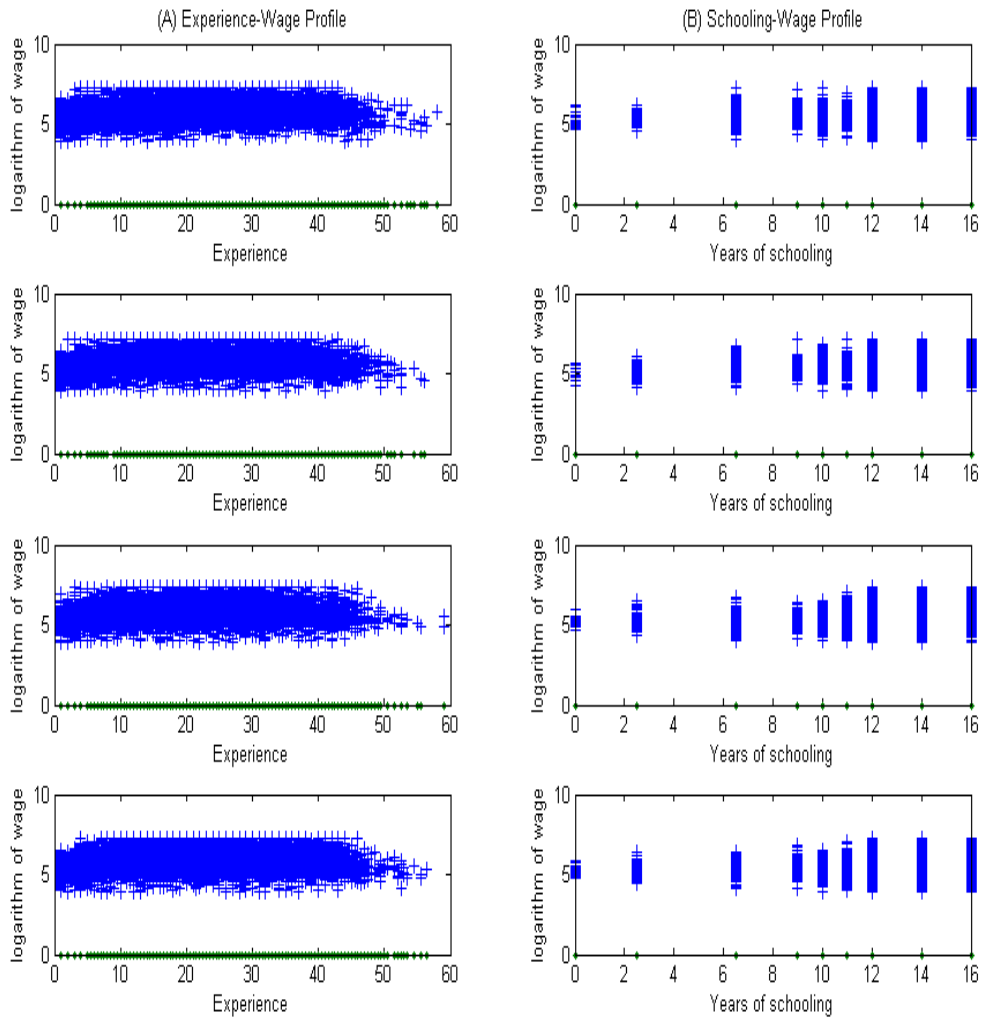


Figure 1: Experience-Wage and Education-Wage profiles. The four columns correspond to years 1990, 1995, 2000, and 2005 from the top to the bottom. The sparsity of the experience variable is also plotted along the experience axis.

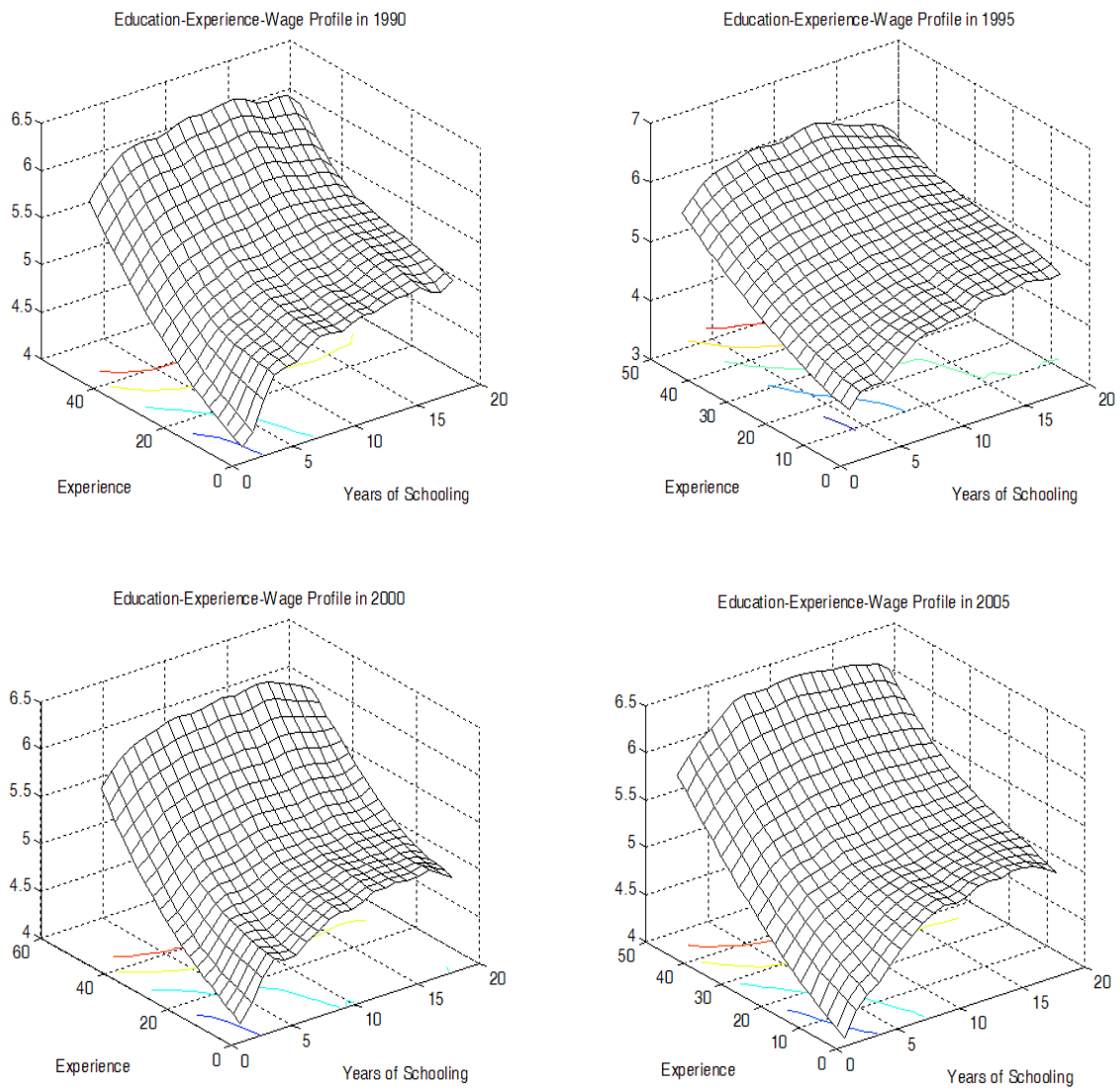


Figure 2: Education-Experience-Wage profile resulting from the partially linear models

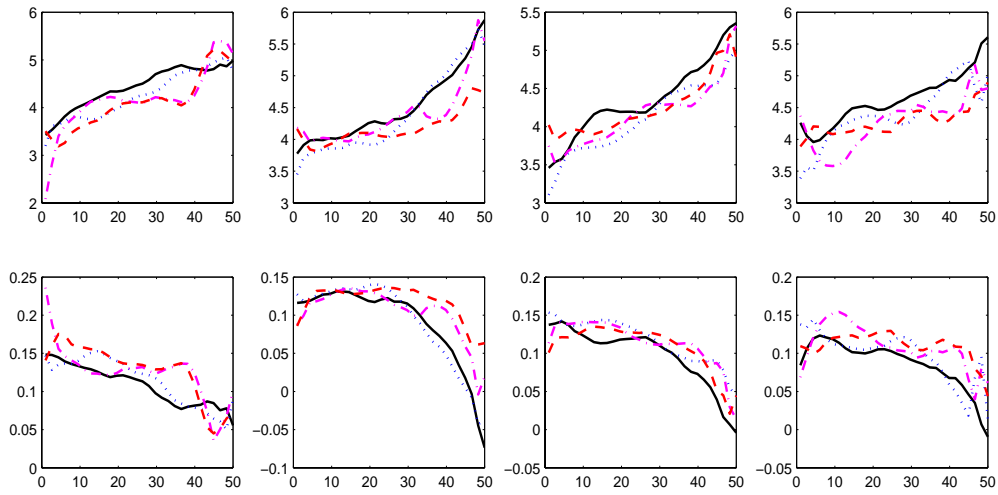


Figure 3: Plots of $a_1(\text{Experience}, \text{Region}, :)$ and $a_2(\text{Experience}, \text{Region}, :)$ averaging over other categorical variables. Horizontal axis: *Experience*. Vertical axis: a_1 or a_2 . The two rows correspond to a_1 and a_2 respectively from the top to the bottom. The four columns correspond to *Region* = Northeast, Midwest, South and West from the left to the right column. 1990: solid line, 1995: dotted line, 2000: dashdot line, 2005: dashed line.

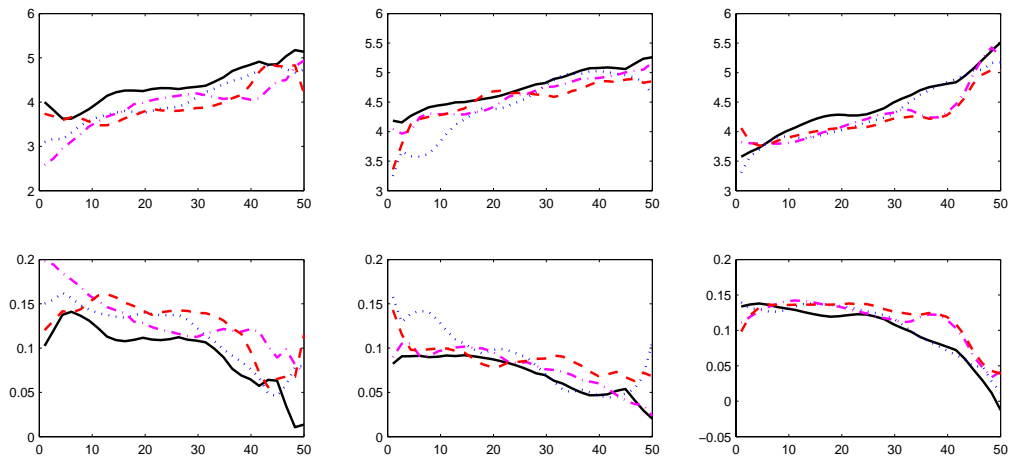


Figure 4: Plots of $a_1(\text{Experience}, \text{Race}, :)$ and $a_2(\text{Experience}, \text{Race}, :)$ averaging over other categorical variables. Horizontal axis: *Experience*. Vertical axis: a_1 or a_2 . The two rows correspond to a_1 and a_2 from the top to the bottom. The three columns correspond to *Race* = Otherwise, Hispanic, and White from the left to the right column. 1990: solid line, 1995: dotted line, 2000: dashdot line, 2005: dashed line.

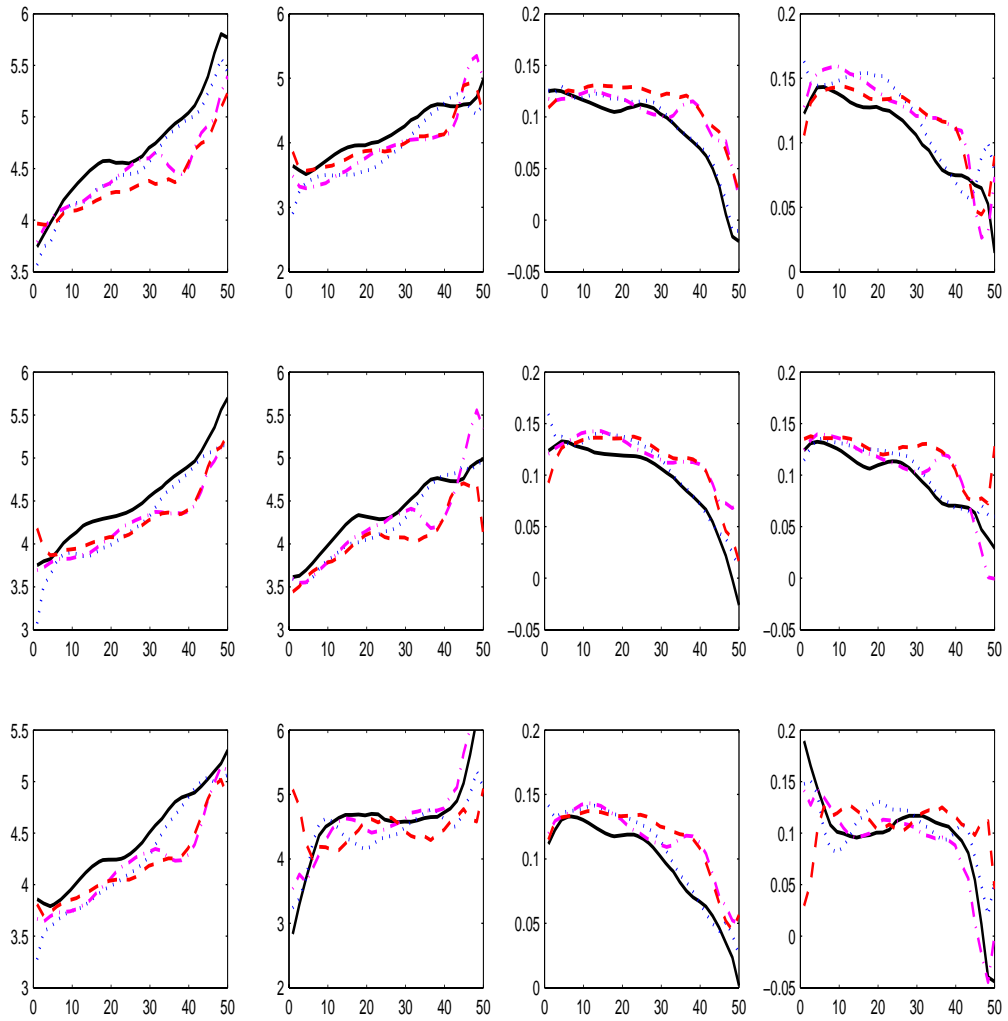


Figure 5: Plots of $a_1(\text{Experience}, \text{Gender}, :)$ and $a_2(\text{Experience}, \text{Gender}, :)$ (1st row), $a_1(\text{Experience}, \text{Single}, :)$ and $a_2(\text{Experience}, \text{Single}, :)$ (2nd row), $a_1(\text{Experience}, \text{Veteran}, :)$ and $a_2(\text{Experience}, \text{Veteran}, :)$ (3rd row), averaging over other categorical variables. Horizontal axis: *Experience*. Vertical axis: a_1 or a_2 . First row: the four columns from the left to the right correspond to a_1 for male, a_1 for female, a_2 for male, and a_2 for female, respectively. Second row: the four columns from the left to the right correspond to a_1 for non-single, a_1 for single, a_2 for non-single, and a_2 for single, respectively. Third row: the four columns from the left to the right correspond to a_1 for non-veteran, a_1 for veteran, a_2 for non-veteran, and a_2 for veteran, respectively. 1990: solid line, 1995: dotted line, 2000: dashdot line, 2005: dashed line.

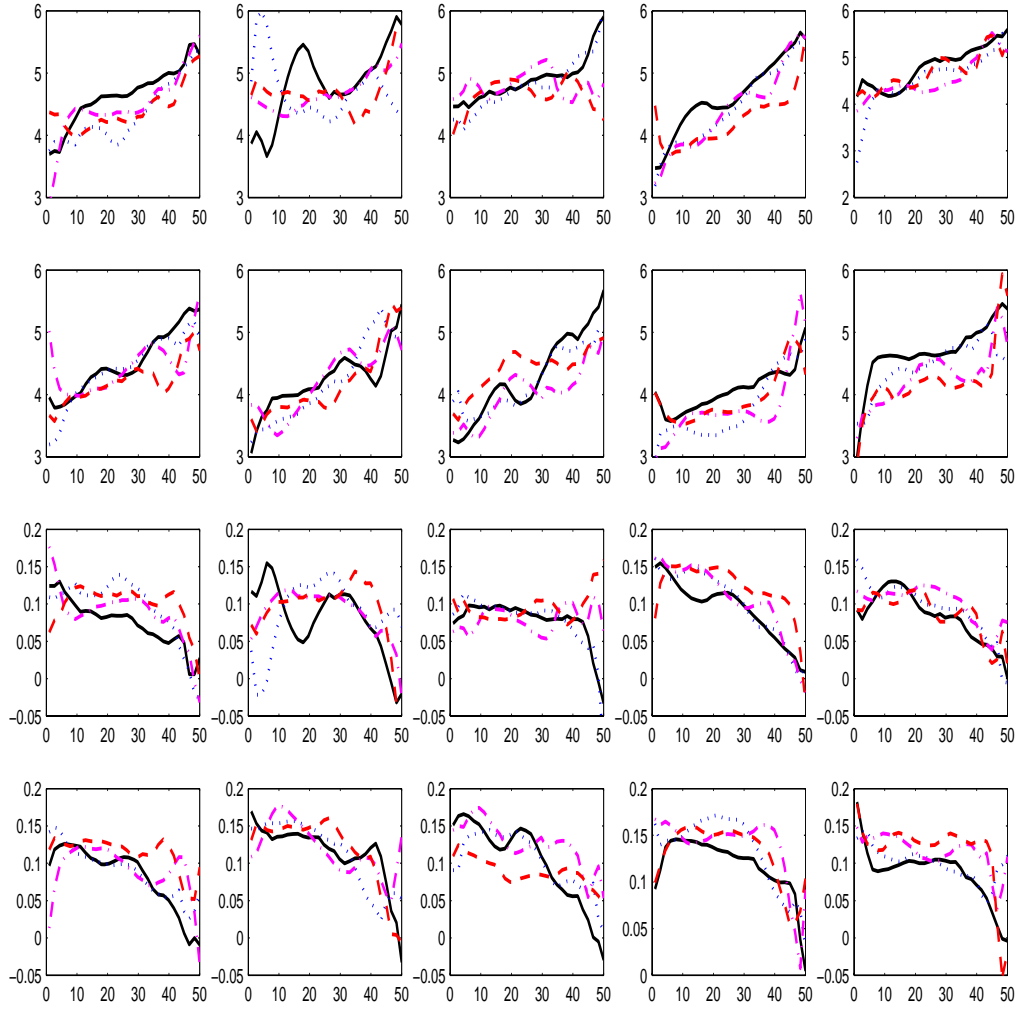


Figure 6: Plots of $a_1(Experience, Industry, :)$ and $a_2(Experience, Industry, :)$ averaging over other categorical variables. Horizontal axis: *Experience*. Vertical axis: a_1 or a_2 . The first two rows correspond to a_1 , and the last two rows correspond to a_2 . For rows 1 and 3, the five columns from the left to the right correspond respectively to *Industry* = Agriculture, Mining, Construction, Manufacturing, and Transportation. For rows 2 and 4, the five columns from the left to the right correspond respectively to *Industry* = Wholesale and return, Finance, Personal services, Professional services, and Public administration. 1990: solid line, 1995: dotted line, 2000: dashdot line, 2005: dashed line.

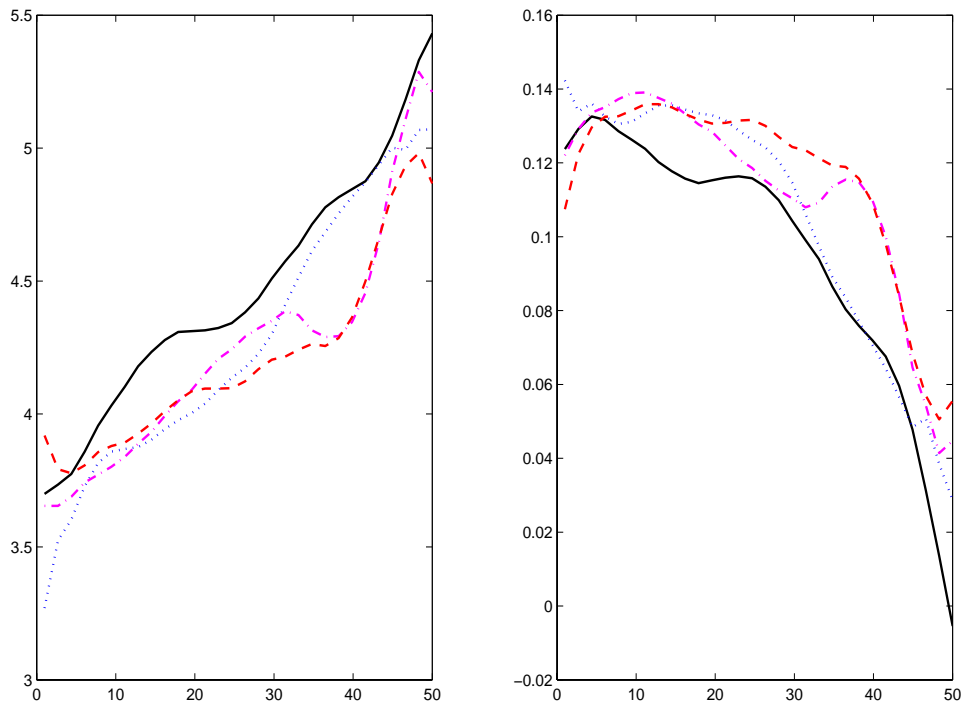


Figure 7: Plots of $a_1(\text{Experience}, :)$ and $a_2(\text{Experience}, :)$ averaging over all categorical variables. Horizontal axis: *Experience*. Vertical axis: a_1 or a_2 . The two columns from the left to the right correspond to a_1 and a_2 , respectively. 1990: solid line, 1995: dotted line, 2000: dashdot line, 2005: dashed line.