Error-in-Variables Jump Regression Using Local Clustering

Yicheng Kang[1], Xiaodong Gong[2], Jiti Gao[3] and Peihua Qiu[4]
JPMorgan[1], University of Canberra[2], Monash University[3] and University of Florida[4]

Abstract

Error-in-Variables (EIV) regression is widely used in econometric models. The statistical analysis becomes challenging when the regression function is discontinuous and the distribution of measurement error is unknown. In this paper, we propose a novel jump-preserving curve estimation method. A major feature of our method is that it can remove the noise effectively while preserving the jumps well, without requiring much prior knowledge about the measurement error distribution. The jump-preserving property is achieved mainly by local clustering. We show that the proposed curve estimator is statistical consistent, and it performs favorably, in comparison with an existing jump-preserving estimator. Finally, we demonstrate our method by an application to a health tax policy study in Australia.

Keywords: Clustering; Demand for private health insurance; Kernel smoothing; Local regression; Measurement errors; Price elasticity.

1 Introduction

This research is motivated by our attempt to study the impact of the Medical Levy Surcharge (MLS) tax policy on the take-up rate of the private health insurance (PHI) in Australia. People in Australia are liable of MLS (which is about 1 percent of their annual taxable incomes) if they do not buy PHI and their annual taxable incomes are above a certain level. For example, the thresholding level for single individuals was $50,000 per annum in the 2003-04 financial year, where the dollar sign "$" used here and throughout the paper represents the Australian Dollar (AUD). The major purpose of MLS was to give people more choices of health insurance and take a certain pressure off the public medical system. Both policy makers and economists are interested in studying the impact of this policy on the relationship between the PHI take-up rate and the annual taxable income. It was expected that this policy would generate a jump in the PHI take-up rate around the thresholding taxable income. This discontinuous relationship could be used to evaluate the impact of the policy. However, such relationship becomes challenging to analyze after the Australian Tax Office (ATO) perturbed the income data

by adding random numbers to them, out of privacy consideration, because the distribution that generates the random numbers was not revealed.

In the literature, jump regression analysis (cf., Qiu 2005) provides a natural framework for studying discontinuous relationship between random variables. In that framework, two approaches have been suggested for estimating a discontinuous curve. The first approach, called the *indirect* approach, estimates the discontinuity locations first and then considers different segments of the design interval, in which the underlying function is assumed to be continuous and can be estimated as usual. See, for example, Eubank and Speckman (1994), Gijbels et al. (1999), Gijbels and Goderniaux (2004), Kang and Qiu (2014), Kang et al. (2015), Muller (1992), Müller (2002), Qiu (1991), Qiu et al. (1991), Qiu and Kang (2015), Wu and Chu (1993), among others. The second approach, called the *direct* approach, estimates the regression curve directly, without first estimating the number and locations of discontinuities. Methods based on this idea include Gijbels et al. (2007), McDonald and Owen (1986), Qiu (2003), and the references therein. Most existing jump-preserving estimation methods assume that the explanatory variable does not have any measurement error involved. Error-in-Variables (EIV) regression models, on the other hand, allow measurement error in the explanatory variables. But most of them assume that the regression function is smooth and that the measurement error distribution is known or it can be estimated reasonably well beforehand (cf., Carroll et al. 1999, 2012, Comte and Taupin 2007, Cook and Stefanski 1994, Delaigle and Meister 2007, Fan and Masry 1992, Fan and Truong 1993, Hall and Meister 2007, Staudenmayer and Ruppert 2004, Stefanski 2000, Stefanski and Cook 1995, and Taupin 2001).

In this paper, we propose a jump-preserving curve estimation method for discontinuous EIV regression models. The proposed method is a *direct* approach without explicitly detecting jumps first and thus it is easy to use. Another feature of our method is that it does not require the measurement error distribution to be specified beforehand, making it applicable to many real problems. The remainder of this article is organized as follows. In Section 2, our proposed method is described in detail. In Section 3, some asymptotic properties of the proposed estimator are discussed. In Section 4, the numerical performance is evaluated by simulated examples. In Section 5, the proposed method is applied to the PHI data. Several remarks conclude the article in Section 6. Some technical details are provided in a supplementary file.

2 Proposed Methodology

Let $\{(W_i, Y_i) : i = 1, \cdots, n\}$ be independent and identically distributed (i.i.d.) observations from the model described below.

$$Y_i = g(X_i) + \varepsilon_i, \tag{1}$$
$$W_i = X_i + \sigma_n U_i, \tag{2}$$

where $i = 1, \cdots, n$, g is the unknown regression function with possible discontinuities, Y_i is the ith observation of the response variable, X_i is the ith observation of the unobservable explanatory variable, ε_i's are i.i.d. random errors with mean 0 and unknown variance $\tau^2 > 0$, W_i is the observed value of X_i with a measurement error, $\sigma_n > 0$ denotes the standard deviation of the measurement error in X_i, and U_i is the standardized measurement error with mean 0 and variance 1. It is also assumed that U_i's are i.i.d., U_i is independent of both X_i and Y_i, the distribution of U_i, denoted as f_U, and the distribution of X_i, denoted as f_X, are both unknown. Without loss of generality, assume that the design interval is $[0, 1]$. Our major goal is to estimate $g(x)$ from the observed data.

Our idea of estimating a regression function with possible jump points of unknown jump locations is that each point in the design interval is a potential jump point and thus the estimation method should adapt at each point to a possible discontinuity. Next, we describe the proposed method in detail. For any given point $x \in [h_n, 1 - h_n]$, where $h_n \in (0, 1/2)$ is a bandwidth parameter, consider a small neighborhood of x defined by

$$N(x; h_n) = \{z \in (0, 1) : |z - x| \le h_n\},$$

and the following local linear kernel (LLK) smoothing procedure:

$$\min_{a,b} \sum_{N(x;h_n)} [Y_i - a - b(W_i - x)]^2 K\left(\frac{W_i - x}{h_n}\right), \tag{3}$$

where K is a density kernel function with support $[-1, 1]$. Let $\left(\widehat{a}_n(x), \widehat{b}_n(x)\right)$ be the solution to (a, b) in (3). Then, the weighted residual mean squares (WRMS) at x is defined by

$$\mathrm{WRMS}_n(x) = \frac{\sum_{N(x;h_n)} \left[Y_i - \widehat{a}_n(x) - \widehat{b}_n(x)(W_i - x)\right]^2 K\left(\frac{W_i - x}{h_n}\right)}{\sum_{N(x;h_n)} K\left(\frac{W_i - x}{h_n}\right)}. \tag{4}$$

If x is a jump point, then the jump structure of the regression function would be dominant even when there is a measurement error involved. This

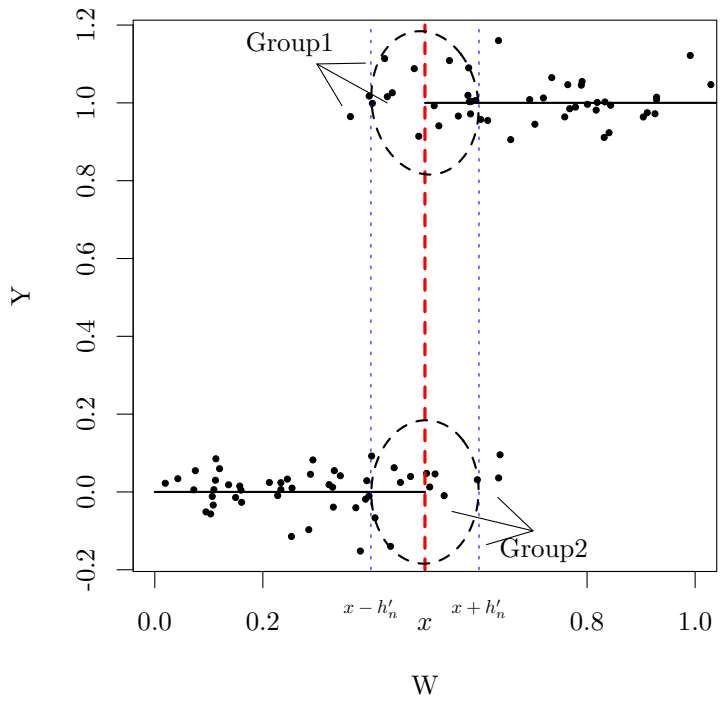

Figure 1: The solid line denotes the regression function $g(\cdot)$ that has a jump at $x = 0.5$ (marked by the vertical dashed line). The dark points denote observations of (W, Y) where W is the observed value of X with measurement error involved. It can be seen that the jump structure of $g(\cdot)$ is quite visible among observations in $N(x; h'_n)$ (i.e., those fall between the two vertical dotted lines) even in the presence of measurement error.

fact is illustrated in Figure 1. By this observation, if x is near a jump point, there should be a significant evidence of lack-of-fit of the LLK smoothing procedure (3). In other words, $\mathrm{WRMS}_n(x)$ would be relatively large. Thus, if the following is true:

$$\mathrm{WRMS}_n(x) > u_n, \tag{5}$$

where u_n is a threshold value, then x is likely to be close to a jump point, and we cannot use all observations near x to estimate $g(x)$ because it would blur the jump otherwise. When there is no measurement error in X, the one-sided estimators can estimate $g(x)$ reasonably well (cf., Qiu 2003). In the case when X has measurement error involved, such estimators are unavailable because X_i's are no longer observable. It may be problematic if we simply replace X_i's by W_i's for constructing a one-sided estimator because we do not know whether a specific value X_i is located on the right (or left) side of x when its observed value W_i is on the right (or left) side of x, due to the measurement error. To overcome this difficulty, we can make use of the fact that the jump structure of $g(\cdot)$ is still quite visible even in the presence of measurement error (cf., Figure 1). We suggest classifying all observations in the neighborhood $N(x; h'_n)$ of x into two significantly separated groups, (i.e., *Group 1* versus *Group 2* in Figure 1), where the bandwidth parameter h'_n could be different from h_n. Then, we can estimate $g(x)$ using the observations in one group. As long as the observations are properly clustered, the jump should be preserved well.

Next, we describe our clustering procedure mentioned above. An ideal classification would put observations whose unobservable X values are on the same side of the jump location into a same group. So, a classification should be reasonable when certain separation measure reaches the maximum. Intuitively, if the two groups of observations are well separated, the within-group variability would be small and the between-group variability would be large. Consequently, the ratio of between-group variability and within-group variability would be large. Therefore, we can use this ratio as a separation measure of the two groups. Specifically, let $G(x; h'_n) = \{(W_i, Y_i) : W_i \in N(x; h'_n)\}$, $G_l(x; h'_n)$ and $G_r(x; h'_n)$ be a partition of $G(x; h'_n)$ (i.e., $G(x; h'_n) = G_l(x; h'_n) \cup G_r(x; h'_n)$ and $G_l(x; h'_n) \cap G_r(x; h'_n) = \emptyset$), and

$$\overline{W}_l = \frac{1}{|G_l(x; h'_n)|} \sum_{(W_i,Y_i)\in G_l(x;h'_n)} W_i, \qquad \overline{Y}_l = \frac{1}{|G_l(x; h'_n)|} \sum_{(W_i,Y_i)\in G_l(x;h'_n)} Y_i,$$

$$\overline{W}_r = \frac{1}{|G_r(x; h'_n)|} \sum_{(W_i,Y_i)\in G_r(x;h'_n)} W_i, \qquad \overline{Y}_r = \frac{1}{|G_r(x; h'_n)|} \sum_{(W_i,Y_i)\in G_r(x;h'_n)} Y_i,$$

where $|A|$ denotes the number of elements in the pointset A. Next, we con-

sider the following LLK smoothing procedures:

$$\min_{a,b} \sum_{(W_i,Y_i)\in G_l(x;h'_n)} [Y_i - a - b(W_i - x)]^2 \, K\left(\frac{W_i - x}{h'_n}\right), \tag{6}$$

and

$$\min_{a,b} \sum_{(W_i,Y_i)\in G_r(x;h'_n)} [Y_i - a - b(W_i - x)]^2 \, K\left(\frac{W_i - x}{h'_n}\right). \tag{7}$$

And the WRMS's defined in (4) can be computed after $N(x;h_n)$ is replaced by $G_l(x;h'_n)$ and $G_r(x;h'_n)$, respectively. They are denoted as $\mathrm{WRMS}_l(x;h'_n)$ and $\mathrm{WRMS}_r(x;h'_n)$. Then, we define the following separation measure:

$$T(G_l(x;h'_n), G_r(x;h'_n)) = \frac{(\overline{W}_l - \overline{W}_r)^2 + (\overline{Y}_l - \overline{Y}_r)^2}{\mathrm{WRMS}_l(x;h'_n) + \mathrm{WRMS}_r(x;h'_n)}. \tag{8}$$

It can be seen that the numerator in (8) represents between-group variability and the denominator represents within-group variability. Let $G^*_l(x;h'_n)$ and $G^*_r(x;h'_n)$ denote the partition that maximizes (8). In practice, solving the optimization in (8) by exhaustive search would be too time-consuming (cf., Everitt et al. 2011, Chapter 5). The more efficient algorithm proposed in Hartigan and Wong (1979) is well received in the literature. We adopt that algorithm in this paper. Let $\widehat{a}_{l,n}(x)$ and $\widehat{a}_{r,n}(x)$ denote the solution to a in (6) and (7), respectively. If $|G^*_l(x;h'_n)| > |G^*_r(x;h'_n)|$, then it is more likely for x to be on the same side of the jump point as the X values of those observations in $G^*_l(x;h'_n)$. So, our proposed estimator of $g(x)$ is

$$\widehat{g}_n(x) = \begin{cases} \widehat{a}_{l,n}(x), & \text{if } |G^*_l(x;h'_n)| > |G^*_r(x;h'_n)|, \\ \widehat{a}_{r,n}(x'), & \text{otherwise.} \end{cases} \tag{9}$$

The proposed jump-preserving curve estimation procedure is summarized as follows.

Jump-preserving Curve Estimation Procedure

Step 1: For any given x, compute its WRMS by (4).

Step 2: If (5) is true, go to Step 3. Otherwise, estimate $g(x)$ by $\widehat{a}_n(x)$, the solution to a in (3).

Step 3: Cluster the observations in $G(x;h'_n)$ by maximizing (8), then compute $\widehat{g}_n(x)$ by (9).

In the proposed estimation procedure (3)–(9), there are three parameters, h_n, h'_n and u_n, to choose. Note that h_n is used for two purposes: to flag possible jump points in (5) and to estimate the curve in continuity regions. The purpose that h_n serves in our procedure is similar to what the bandwidth parameter does in the conventional kernel smoothing. Thus, we suggest to select h_n to minimize the following leave-one-out cross validation score:

$$\min_{h_n} \sum_{i=1}^{n} \left[Y_i - \widehat{a}_n^{(-i)}(W_i) \right]^2,$$

where $\widehat{a}_n^{(-i)}(\cdot)$ denotes the estimate $\widehat{a}_n(\cdot)$ when the ith observation (W_i, Y_i) is omitted.

Next, we discuss the selection of h'_n and u_n. In simulation studies, the true regression function g could be known. Then, once h_n is selected, (h'_n, u_n) can be chosen to be the pair that minimizes the Mean Square Error (MSE), defined as

$$\text{MSE}\left(\widehat{g}, g; h'_n, u_n\right) = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{g}(x_i) - g(x_i) \right]^2, \tag{10}$$

where $\{x_1, x_2, \cdots, x_n\}$ are equally spaced values on $[0, 1]$. In practice, g is usually unknown. In such cases, we suggest the following bootstrap selection procedure:

- For a given bandwidth value $h'_n > 0$ and threshold value $u_n > 0$, apply the proposed estimation procedure (3)–(9) to the original dataset $\{(W_1, Y_1), (W_2, Y_2), \cdots, (W_n, Y_n)\}$, and obtain an estimator of g, denoted as $\widehat{g}(\cdot\,; h'_n, u_n)$.

- Draw with replacement n times from the original dataset to obtain the first bootstrap sample, denoted as $\{(\widetilde{W}_1^{(1)}, \widetilde{Y}_1^{(1)}), (\widetilde{W}_2^{(1)}, \widetilde{Y}_2^{(1)}), \cdots, (\widetilde{W}_n^{(1)}, \widetilde{Y}_n^{(1)})\}$.

- Apply the proposed estimation procedure (3)–(9) to the first bootstrap sample, and obtain the first bootstrap estimator of g, denoted as $\widetilde{g}^{(1)}(\cdot\,; h'_n, u_n)$.

- Repeat the previous two steps B times and obtain B bootstrap estimators of g: $\{\widetilde{g}^{(1)}(\cdot\,; h'_n, u_n), \widetilde{g}^{(2)}(\cdot\,; h'_n, u_n), \cdots, \widetilde{g}^{(B)}(\cdot\,; h'_n, u_n)\}$.

- Then, the bandwidth h'_n and the threshold u_n are chosen to be the minimizer of

$$\min_{h'_n, u_n} \frac{1}{B} \sum_{k=1}^{B} \frac{1}{n} \sum_{i=1}^{n} \left[\widetilde{g}^{(k)}(x_i; h'_n, u_n) - \widehat{g}(x_i; h'_n, u_n) \right]^2. \tag{11}$$

3 Asymptotic Properties

In this section, we discuss some asymptotic properties of the proposed estimation procedure (3)–(9). To this end, we have the theorem below.

Theorem 1. *Suppose that the following conditions hold:*

(1) $\{(W_1, Y_1), (W_2, Y_2), \cdots, (W_n, Y_n)\}$ *are i.i.d. observations from models* (1) *and* (2).

(2) $g(\cdot)$ *is a bounded, piecewise continuous function defined on* $[0, 1]$ *with finitely many jump points in* $[0, 1]$*; at each jump point,* $g(\cdot)$ *has finite one-sided limit; its first-order derivative,* $g'(\cdot)$*, is also a bounded function and is continuous on* $[0, 1]$ *except on those jump points; at each jump point,* $g'(\cdot)$ *also has finite one-sided limit. Denote the set of all the jump points by* S.

(3) The support of f_X *is* $[0, 1]$*;* f_X *is uniformly continuous, bounded, and positive on* $(0, 1)$ *and has bounded derivatives on* $(0, 1)$.

(4) f_U *is continuous on its support, symmetric about* 0 *with* $f_U(0) > 0$ *and satisfies the conditions that* $\int_{-\infty}^{\infty} u f_U(u)\, du = 0$ *and* $\int_{-\infty}^{\infty} u^2 f_U(u)\, du = 1$.

(5) $E|\varepsilon_1|^4 < \infty$.

(6) $h_n = o(1)$, $1/(n^{1/3} h_n) = o(1)$, $h'_n = o(1)$, $\sigma_n^2/h'_n = o(1)$, *and* $1/(n^{1/3} h'_n) = o(1)$

(7) The kernel function K *is a Lipschitz-1 continuous density function with support* $[-1, 1]$ *and is symmetric about* 0.

(8) $u_n = \tau^2 + \delta_n$*, where* δ_n *is sequence of positive numbers such that* $\delta_n = o(1)$ *and that* $[h_n^2 + \sigma_n^2 + (\log n)^{1+\gamma}/(nh_n)^{\beta}]/\delta_n = o(1)$ *for some* $\gamma > 0$ *and some* $\beta \in (0, 1/4)$.

Then, we have, with probability 1,

$$\widehat{g}_n(x) - g(x) = \begin{cases} O\left(h_n^2 + \sigma_n^2 + (\log n)^{1+\gamma}/(nh_n)^{\beta}\right), & \text{if } d_E(x, S) > h_n, \\ O\left(h_n'^2 + \sigma_n^2 + (\log n)^{1+\gamma}/(nh'_n)^{\beta}\right), & \text{otherwise,} \end{cases}$$

where S *is the set of all true jump points and* $d_E(x, S) = \min_{x_s \in S} |x - x_s|$.

The Theorem 1 shows that the proposed estimation procedure (3)–(9) estimates $g(\cdot)$ consistently under some regularity conditions. Its proof is given in a supplementary file.

Remark 1: Theorem 1 requires that the measurement error standard deviation σ_n tends to 0 when the sample size increases. In the literature, it has been pointed out that this condition is needed for consistently estimating the regression function when its observations have measurement errors involved and when little prior information about the measurement error distribution is available (cf., Delaigle 2008).

Remark 2: The rate of convergence of σ_n to 0 does not need to be comparable to that of h_n for our proposed method to flag all jump points correctly (see the supplementary file for details). However, the condition $\sigma_n^2/h_n' = o(1)$ is required for classifying a jump point into the correct cluster. This is a weaker condition than the one $\sigma_n/h_n' = o(1)$ which would be required by existing jump regression methods (e.g., Gijbels et al. 2007, Qiu 2003) to ensure consistency.

4 Numerical Studies

In this section, we study the numerical performance of the proposed method described in Section 2, which are organized in two subsections. Section 4.1 presents some simulation examples related to the procedure (3) – (11). Section 4.2 compares the proposed curve estimator to the piecewise-linear kernel estimator (PLKE) that ignores the measurement error (see Qiu 2003 for a detailed discussion of PLKE).

4.1 Numerical Performance of the Proposed Methodology

In this subsection, the performance of the proposed estimation procedure is evaluated using the following two true regression functions:

$$g_1(x) = (3x^2 + 0.53)\mathbb{1}_{\{0.3\le x<0.7\}} + (2x^2 + 2.22)\mathbb{1}_{\{0.7\le x\le 1\}},$$
$$g_2(x) = \cos(4\pi(0.5 - x))\mathbb{1}_{\{0\le x<0.5\}} - \cos((4\pi(x - 0.5))\mathbb{1}_{\{0.5\le x\le 1\}},$$

where $\mathbb{1}_{\{.\}}$ is an indicator function taking the value of 1 if the argument in the brace is true and 0 otherwise. g_1 and g_2 are graphed in Figure 2(a) and Figure 2(c), respectively. It can be seen that g_1 has two jump points. One is at $x = 0.3$ with jump size 0.8 and the other one is at $x = 0.7$ with jump size 0.8. g_2 has a single jump of size 2 at $x = 0.5$. For each regression function, we consider cases when the sample size n equals 500 or 1000, $f_X \sim \text{Unif}[0, 1]$, and f_U is either a Normal, a Laplace, or a Uniform distribution with $\text{E}(U) = 0$ and $\text{Var}(U) = 1$. τ and σ_n are fixed at 0.1 and 0.05, respectively. In each combination of g, n, and f_U, the simulation is repeated 200 times. For each

given bandwidth h'_n and threshold value u_n, 200 values of $\text{MSE}(\widehat{g}, g; h'_n, u_n)$ defined in (10) are computed. Their average is called the Average Mean Square Error (AMSE) and is denoted as $\text{AMSE}(h'_n, u_n)$. The minimizer of $\text{AMSE}(h'_n, u_n)$ is called the optimal bandwidth and the optimal threshold, and is denoted as h'_{opt} and u_{opt}, respectively. We also compute the bandwidth value and threshold value using the proposed bootstrap selection procedure. Such bandwidth and threshold are called the bootstrap bandwidth and the bootstrap threshold, denoted as h'_{bt} and u_{bt}, respectively. Throughout this section, if there is no further specification, the bootstrap sample size B is chosen to be 200, and K used in (3), (4), (6), and (7) is chosen to be the Epanechnikov kernel function (i.e., $K(x) = 0.75(1-x^2)\mathbb{1}_{\{|x|\leq 1\}}$). The values of h'_{opt}, u_{opt}, h'_{bt}, u_{bt}, and $\text{AMSE}(h'_{opt}, u_{opt})$ are presented in Table 1.

Table 1: Numerical summary of two simulation examples based on 200 replicated simulations

g	n	f_U	h'_{opt}	h'_{bt}	u_{opt}	u_{bt}	AMSE
g_1	500	Normal	0.06	0.06	0.09	0.10	0.0229
		Laplace	0.05	0.07	0.09	0.11	0.0220
		Uniform	0.08	0.09	0.07	0.07	0.0237
	1000	Normal	0.07	0.06	0.08	0.10	0.0155
		Laplace	0.04	0.06	0.09	0.11	0.0158
		Uniform	0.09	0.08	0.07	0.07	0.0175
g_2	500	Normal	0.04	0.05	0.37	0.39	0.0675
		Laplace	0.05	0.03	0.32	0.34	0.0572
		Uniform	0.06	0.08	0.39	0.41	0.0779
	1000	Normal	0.04	0.04	0.34	0.36	0.0521
		Laplace	0.04	0.03	0.29	0.31	0.0425
		Uniform	0.05	0.05	0.38	0.40	0.0589

From the table, it can be seen that (i) the performance of the proposed estimation procedure improves as the sample size n increases, and (ii) the bootstrap selection procedure chooses parameters close to the optimal ones.

Next, in the case when $n = 500$ and f_U is Normal, the realizations of $\{(W_i, Y_i) : i = 1, \cdots, n\}$ when g is g_1 and g_2 and their corresponding estimates are shown in Figure 2(b) and Figure 2(d), respectively. It can be seen

that our curve estimation procedure preserves the jumps well in the presence of measurement error. It also can be seen that there were little kinks near the jump points in the estimated curves. This is because when $\mathrm{WRMS}_n(x)$ approaches to but does not exceed the threshold u_n, $\widehat{g}_n(x) = \widehat{a}_n(x)$ which uses observations from both sides of the jump point. As x gets closer to the jump point, once $\mathrm{WRMS}_n(x)$ exceeds u_n, the clustering procedure kicks in and $\widehat{g}_n(x)$ will use the observations from one cluster only. Thus, the slight kinky behavior is mainly caused by the hard thresholding used in the proposed procedure.

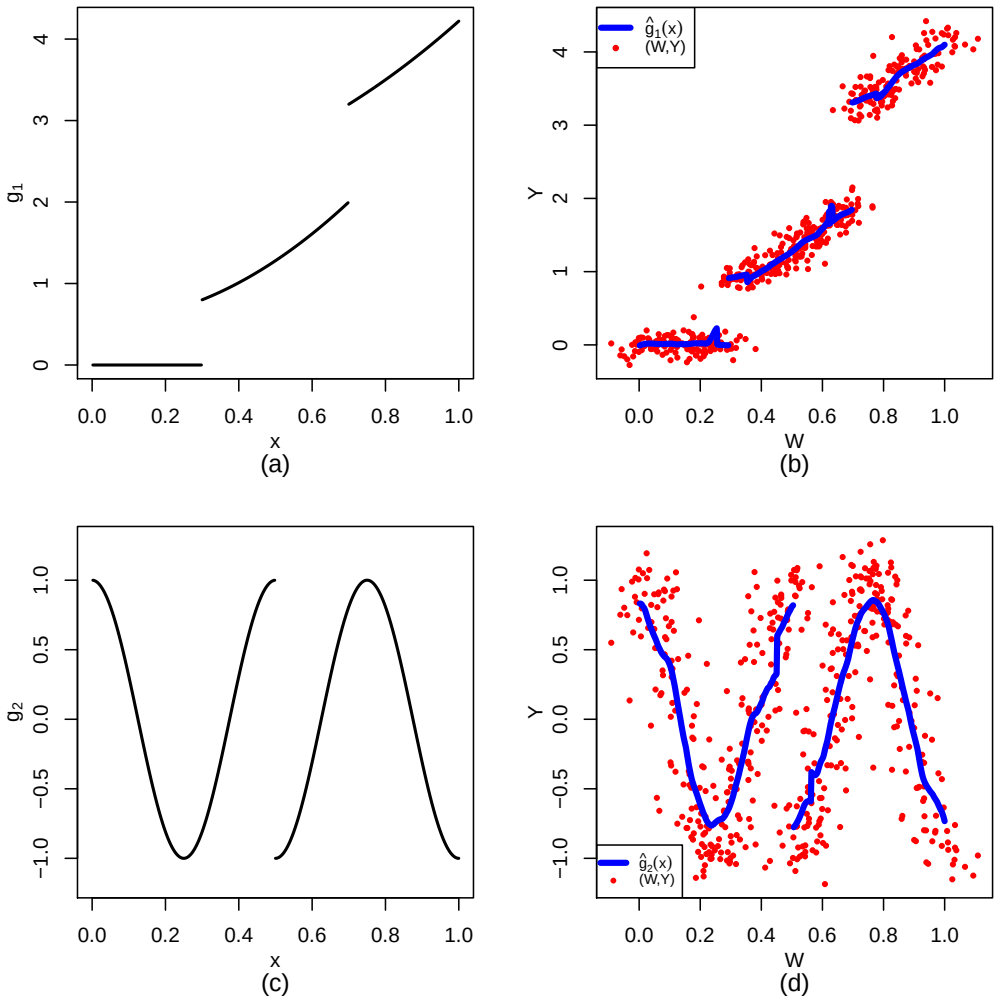

Figure 2: (a): g_1. (b): One realization with the regression function g_1 (dark dots) and $\widehat{g}_1$ (solid line). (c): g_2. (d): One realization with the regression function g_2 (dark dots) and $\widehat{g}_2$ (solid line).

4.2 Comparison to the PLKE Procedure

The PLKE proposed by Qiu (2003) is a *direct* curve estimation method that preserves jumps well when there is no measurement error involved. In this subsection, we compare our proposed procedure with PLKE procedure in an artificial example. The proposed procedure is denoted as NEW and the PLKE procedure is denoted as PLKE. Assume that the regression function is

$$g(x) = \begin{cases} -\frac{25}{9}(x-0.6)^2, & \text{if } x \in [0, 0.6), \\ 4(x-0.6)^3 + 0.5, & \text{if } x \in [0.6, 1.0]. \end{cases}$$

It can be seen that $g(x)$ is a piecewise polynomial with a jump size 0.5 at $x = 0.6$, as plotted in Figure 3(a) (the solid line). In this numerical comparison, we choose the sample size n to be 500, τ to be 0.05, f_U to be $N(0, 0.1^2)$, and f_X to be either Unif[0, 1], Beta(2,2), Beta(3,2) or Beta(2,3). In each case, the simulation is repeated 200 times, the optimal parameters (i.e., the bandwidth-threshold pair for NEW and the bandwidth parameter for PLKE) are selected based on the AMSE from 200 replicated simulations. The AMSEs and their standard deviations (denoted as SDAMSE) are computed. These results are presented in Table 2. From Table 2, it can be seen that the proposed procedure outperforms the PLKE procedure, across all difference choices of f_X.

Table 2: Numerical comparison of the proposed method NEW with the PLKE method based on 200 replicated simulations. The numbers are in 10^{-3}.

	NEW		PLKE	
f_X	AMSE	SDAMSE	AMSE	SDAMSE
Unif[0,1]	7.2116	0.1938	7.4668	0.1079
Beta(2,2)	14.2586	0.2717	20.4528	2.1102
Beta(3,2)	19.8202	0.5693	42.3686	4.3060
Beta(2,3)	23.2024	3.7319	25.5297	1.1948

Next, one realization of $\{(W_i, Y_i), i = 1, \cdots, n\}$ when f_X is Unif[0, 1] is shown in Figure 3(a). The fitted curve by the proposed procedure and the one fitted by PLKE are shown together in Figure 3(b). It can be seen that PLKE blurred the jump due to the impact of the measurement error whereas the proposed procedure preserves the jump well.

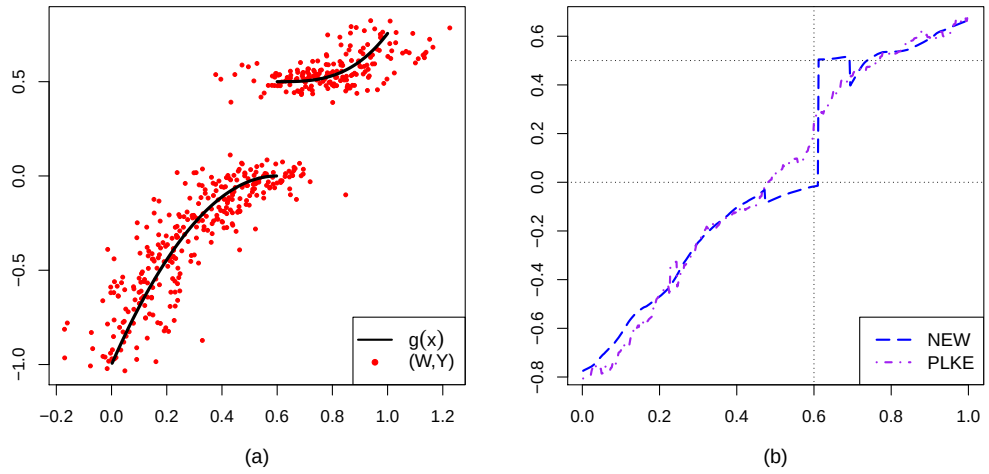

Figure 3: (a): $g(x)$ (solid line) and one realization of $\{(W_i, Y_i), i = 1, \cdots, n\}$ (dark dots). (b): Estimated curves by NEW (long-dashed line) and by PLKE (dot-dashed line), respectively. The dotted lines mark the jump location (i.e., $x_s = 0.6$) and the one-sided limits $\lim_{x \to x_s \pm} g(x)$.

5 Analysis of the PHI Data

In this section, we apply our proposed jump detector to the PHI data for evaluating the impact of MLS on the take-up rate of PHI, as discussed in Section 1. The purposes of introducing PHI in Australia were to give consumers more choices and take some pressure off the public medical system. However, the PHI take-up rate by Australians was very low at the beginning when the PHI was first introduced in 1984, and the take-up rate has been in declining until the end of 1990s (the take-up rate was only about 31 percent at that time) when a series of policies (including MLS) were introduced. Impact of some of these policy measures (e.g., Lifetime Health Cover) has been studied in a few studies, including Butler (2002), Frech et al. (2003), Palangkaraya and Yong (2005), and Palangkaraya et al. (2009). But the role of MLS has not been identified separately yet. The MLS was imposed in 1997 on high-income taxpayers who did not have private insurances. Between 1997-1998 and 2007-2008, the threshold of annual taxable income at which MLS was payable was \$50,000 for singles without children, and \$100,000 for couples. For each dependent child in the household, the threshold increased by \$3,000. So, people having children may lead to multiple jumps in the current PHI data. Unfortunately, we do not have information on the number of children in a family. Also, multiple jump locations within a relatively narrow range would be difficult to distinguish, given the measurement error involved in the PHI data. To mitigate the effect of multiple jumps due to people having

children, this paper focuses only on singles in the current PHI data.

The data used here are from a confidentialised "1% Sample Unit Record File of Individual Income Tax Returns" for the 2003-04 financial year, that was developed by ATO for research purposes. The file contains just over 109,000 records of individual tax returns and detailed information on income from various sources; different types of tax deductions; taxable income; and the take-up of PHI by the individuals. It also contains a limited number of demographic variables, including gender, age group, and marital status. In this paper, we focus on singles between 20 and 69 years old, who were all subject to the same income threshold of $50,000 for the MLS. Therefore, the PHI take-up rate is expected to have a jump around that level of the annual taxable income. In the tax and transfer system or in the health insurance premium regime in Australia, there is no other differential treatment related to the PHI take-up. Other demographic covariates (such as gender and age) would not generate discontinuity in the take-up rate either. So, in the current PHI data, MLS seems to be the only factor responsible for the jump in the take-up rate.

As a method of confidentialisation, ATO 'perturbed' the income variables and the deductions, and provided the following information on the way the data was perturbed: several random numbers within a specified range for each individual were generated, which were converted into a rate (equal probability of being positive or negative) and which was then applied to the various components of the tax return. These rates were applied to the components in a way to try to maintain relationships with similar items. This was achieved by grouping the components into three broad categories: work or employment related income and deductions; investment income and deductions; and business and other income and deductions. Thus, there is some information about the measurement errors in the income data, but the actual distribution of the measurement errors is impossible to be identified based on the provided information. The sample was further restricted to minimize the number of income sources/deduction sources so that the distribution of the error term could be more homogeneous, according to the following criteria: 1) Only those who had positive earnings as the only sources of income were selected; 2) Individuals whose taxable income was not positive (which means their total tax deductions were not less than their earnings) were dropped; and 3) We further dropped individuals whose non-work related deductions formed a significant part of their taxable income—specifically, we dropped those individuals whose work related deductions were less than 90 percent of earnings when the total deductions were more than 10 percent of earnings; whose total deductions were over 50 percent of earnings; or whose total deductions were all non-work related and the total deductions were over 10 percent of their earnings.

The final sample for analysis contains 9,685 records of individual tax returns. By a preliminary analysis, we found that about 26% singles bought PHI in 2003-04, and the PHI take-up rates for those whose annual taxable incomes were below \$50,000 and those whose annual taxable incomes were above that level were quite different. The PHI take-up rate for the former group was about 21%, and it was about 57% for the latter group. Because ATO perturbed the income data by multiplying each original income observation by a random number, we used the income variable in *log* scale in our analysis, so that the additive measurement error assumption in (2) is valid here. Also, the response variable is 0 when an individual did not purchase PHI in 2003-04 and 1 otherwise. We transformed binned observations to meet the model assumption in (1) that the response variable is continuous numerical. Specifically, the bin size is chosen to be 40. For each bin, the average of annual taxable income on *log* scale in that bin is used as the value for the new explanatory variable. And the log odds of the PHI take-up rate (i.e., $\log(p(x)/(1-p(x))$, where $p(x)$ denotes the PHI take-up rate when the bin average of the logarithm of the annual taxable income is x) is used as the transformed response variable. The log odds in the ith bin is computed by

$$\log\left(\frac{N_i + c}{m - N_i + c}\right), \tag{12}$$

where $m = 40$ is the bin size, N_i denotes the number of people in the ith bin who purchased PHI during 2003-04, and c is some positive constant to avoid the numerical instabilities in computing the log odds. (12) is known as the empirical logistic transformation when $c = 0.5$ and it yields some good statistical properties (see Cox 1970 for a detailed discussion). This choice for c is also adopted here. The transformed PHI data is shown in Figure 4 (dark dots). From the figure, it can be seen that there is an abrupt change in the log odds of PHI take-up rate within $[10.75, 11.25]$ (i.e., the annual income is within [\$36315, \$59874]). The impact of the measurement error is also visible.

We then apply our proposed estimation procedure (3)–(9) to the transformed PHI data. The bandwidth and the threshold are chosen to be 0.135 and 0.13, respectively. The results are shown in Figure 4 (dashed line). From the plot, the abrupt change in the log odds is estimated to be at 10.99 ($\approx$ \$59,278). This finding confirms our intuition that people usually act later than they are hit by the MLS. From Figure 4, it can also be seen that the jump size is around 0.4945 in log odds ($\approx$ 12.3% in the PHI take-up rate). This number shows that the impact of the MLS tax policy is quite substantial. For individuals with only one income source, the policy can be considered locally exogenous because the observations to the left and right of (but close to) the jump position are more or less homogeneous except the policy treatment. It

implies that, among the individuals whose annual taxable income is around \$59,278, MLS brings about an extra 12.3% of them onto the private health system. This also implies a negative price elasticity of PHI demand since the jump in the take-up rate can be seen as a response to a price discount in the premium.

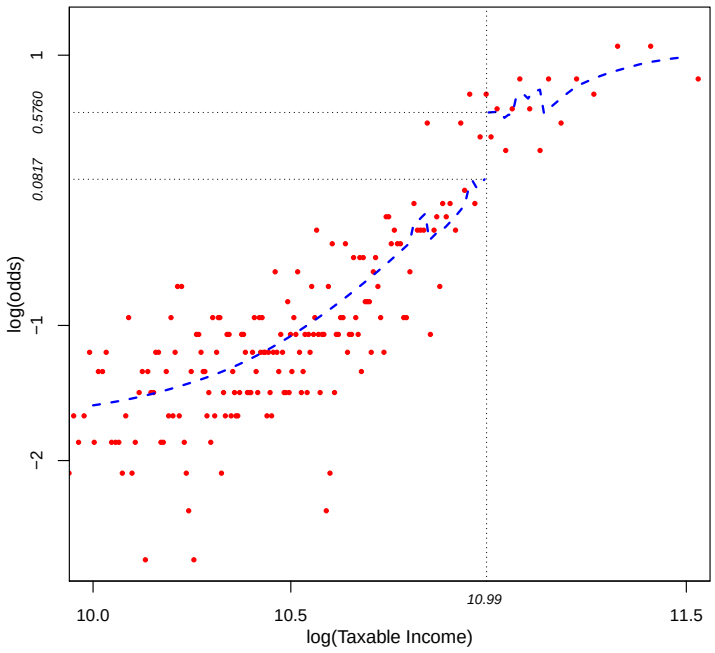

Figure 4: The estimated $\log(p(\cdot)/(1-p(\cdot)))$ (dashed line) and the observations of the PHI data after transformation (solid dots).

6 Concluding Remarks

We have proposed a jump-preserving curve estimation method when the explanatory variable has measurement error involved. A major feature of the proposed method is that it preserves jumps well without requiring much prior knowledge on the measurement error distribution, making it applicable in practice. The challenge caused by measurement error with an unknown distribution is handled by locally clustering of observations by maximizing a separation measure. Also, the proposed method is a direct approach without explicitly detecting the jump points beforehand. Thus, it is easy to use.

There is much room for further improvement of the current method. First, estimated curve by the proposed method exhibits kinky behavior near jump points due to the hard thresholding. Some post-processing modifications may help improve the fitted curve. Second, the regression function at continuity points is estimated by the conventional local linear kernel smoothing procedure and the measurement error is ignored in such cases. It might be possible to estimate the distribution of the measurement error to some extent by making use of the jump structure of the regression function and then refine our estimate of the regression function in continuity regions. Fourth, the proposed bootstrap parameter selection procedure is evaluated by numerical studies only. It requires future research to derive the theoretical justification of its asymptotic behavior.

References

Butler, J. R. (2002). Policy change and private health insurance: Did the cheapest policy do the trick? *Australian Health Review*, 25(6):33–41.

Carroll, R. J., Maca, J. D., and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika*, 86(3):541–554.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2012). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Comte, F. and Taupin, M.-L. (2007). Nonparametric estimation of the regression function in an errors-in-variables model. *Statistica Sinica*, 17:1065–1090.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.

Cox, D. (1970). *The Analysis of Binary Data*. Chapman and Hall, London.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statistica Sinica*, 18(3):1025–1045.

Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association*, 102(480):1416–1426.

Eubank, R. and Speckman, P. (1994). Nonparametric estimation of functions with jump discontinuities. *Lecture Notes-Monograph Series*, pages 130–144.

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. John Wiley and Sons, Ltd., 5th edition.

Fan, J. and Masry, E. (1992). Multivariate regression estimation with errors-in-variables: asymptotic normality for mixing processes. *Journal of multivariate analysis*, 43(2):237–271.

Fan, J. and Truong, Y. K. (1993). Nonparametric regression with errors in variables. *The Annals of Statistics*, 21(4):1900–1925.

Frech, H. E., Hopkins, S., and MacDonald, G. (2003). The australian private health insurance boom: was it subsidies or liberalised regulation? *Economic Papers: A journal of applied economics and policy*, 22(1):58–64.

Gijbels, I. and Goderniaux, A.-C. (2004). Bandwidth selection for change-point estimation in nonparametric regression. *Technometrics*, 46(1):76–86.

Gijbels, I., Hall, P., and Kneip, A. (1999). On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics*, 51(2):231–251.

Gijbels, I., Lambert, A., and Qiu, P. (2007). Jump-preserving regression and smoothing using local linear fitting: a compromise. *Annals of the Institute of Statistical Mathematics*, 59(2):235–272.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *The Annals of Statistics*, 35(4):1535–1558.

Hartigan, J. and Wong, M. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Kang, Y., Gong, X., Gao, J., and Qiu, P. (2015). Jump detection in generalized error-in-variables regression with an application to australian health tax policies. *The Annals of Applied Statistics*, 9:883–900.

Kang, Y. and Qiu, P. (2014). Jump detection in blurred regression surfaces. *Technometrics*, 56:539–550.

McDonald, J. A. and Owen, A. B. (1986). Smoothing with split linear fits. *Technometrics*, 28(3):195–208.

Müller, C. H. (2002). Robust estimators for estimating discontinuous functions. *Metrika*, 55:99–109.

Muller, H.-G. (1992). Change-points in nonparametric regression analysis. *The Annals of Statistics*, 20(2):737–761.

Palangkaraya, A. and Yong, J. (2005). Effects of recent carrot-and-stick policy initiatives on private health insurance coverage in australia. *Economic Record*, 81(254):262–272.

Palangkaraya, A., Yong, J., Webster, E., and Dawkins, P. (2009). The income distributive implications of recent private health insurance policy reforms in australia. *The European Journal of Health Economics*, 10(2):135–148.

Qiu, P. (1991). Estimation of a kind of jump regression functions. *Journal of Systems Science and Complexity*, 4:1–13.

Qiu, P. (2003). A jump-preserving curve fitting procedure based on local piecewise-linear kernel estimation. *Journal of Nonparametric Statistics*, 15(4-5):437–453.

Qiu, P. (2005). *Image Processing and Jump Regression Analysis*. John Wiley & Sons.

Qiu, P., Asano, C., and Li, X. (1991). Estimation of jump regression function. *Bulletin of Informatics and Cybernetics*, 24:197–212.

Qiu, P. and Kang, Y. (2015). Blind image deblurring using jump regression analysis. *Statistica Sinica*, 25:879–899.

Staudenmayer, J. and Ruppert, D. (2004). Local polynomial regression and simulation–extrapolation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):17–30.

Stefanski, L. (2000). Measurement error models. *Journal of the American Statistical Association*, 95(452):1353–1358.

Stefanski, L. and Cook, J. (1995). Simulation-extrapolation: the measurement error jackknife. *Journal of the American Statistical Association*, 90(432):1247–1256.

Taupin, M.-L. (2001). Semi-parametric estimation in the nonlinear structural errors-in-variables model. *Annals of Statistics*, 29(1):66–93.

Wu, J. and Chu, C. (1993). Kernel-type estimators of jump points and values of a regression function. *The Annals of Statistics*, 21(3):1545–1566.