MEASUREMENT ERRORS REGRESSION: A NONPARAMETRIC APPROACH

Jianqing Fan

Department of Statistics
University of North Carolina
Chapel Hill, N.C. 27599

Young K. Truong

Department of Biostatistics
University of North Carolina
Chapel Hill, N.C. 27599

Yonghua Wang

Department of Statistics & Biostatistics
University of North Carolina
Chapel Hill, N.C. 27599

October 25, 1990

Abstract

Nonparametric regression provides a useful tool for exploring the association between the responses and covariates. In many applications, the actual values of the covariates are not known, but are measured with errors or through surrogates. In this paper, we present nonparametric regression techniques to study the association between the response and the underlying unobserved covariate. Deconvolution techniques are used to account for the errors in the covariates. The proposed regression estimators are shown to have limiting normal distributions and methods for constructing confidence intervals are also introduced. Various simulated examples based on both continuous and binary responses are included to illustrate the usefulness of the proposed methods.

---

[0] *Abbreviated title.* Measurement errors regression

AMS 1980 subject classification. Primary 62G20. Secondary 62G05, 62J99.

Key words and phrases. Binary response; Confidence interval; Deconvolution; Errors-in-variables; Kernel estimator; Nonparametric regression.

1 Introduction

Recently, there has been a great deal of interest in the regression problem involving errors-in-covariates. This is due largely to important medical and epidemiologic studies where risk factors are usually available partially. Prentice (1986) presented a study about the effect of radiation on chromosomal aberrations based on 649 survivors of the Hiroshima atomic blast, the amount of radiation actually received is not known but its measurements were obtained through physical models of radioactive transmission along with interview data for each of the survivors, primarily regarding their distance from the blast. Whittemore and Keller (1988) considered the respiratory illness verses exposure to nitrogen dioxide (NO_2) based on group of school children, the actual amount of (NO_2) is not known but was measured by the concentration in the child's room, a surrogate for personal exposure to (NO_2).

To describe this regression problem more formally, consider a response Y and a covariate X°, that is observed through $X = X^\circ + \varepsilon$, where ε is independent of X° and Y. Suppose now the error distribution is known and that it is desired to estimate the regression function $m(x) = E(Y|X^\circ = x)$ based on a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y).

In parametric approach, the regression function is assumed to be a specific function with unknown parameters which are then estimated by (say) maximum likelihood method. See, for example, Armstrong (1985), Stefanski (1985), Stefanski and Carroll (1985, 1987), Prentice (1986), Whittemore and Keller (1988) and Fuller (1987), and Whittemore (1989). One drawback in this approach is that the resulting maximum likelihood equations are usually very complicated. The other is that erroneous conclusions can be made if the parametric model is misspecified. To overcome these problems and to enhance the flexibility in fitting data, we propose a modified kernel method for estimating the regression function in which the idea of deconvolution is involved to account for the effect of errors-in-covariates. As in nonparametric curve estimation, this method does not require a *particular functional form of the regression curve.* Thus it has the advantage for exploring the relationship

between the response and the covariate. Further, it also provides a useful diagnostic tool for regression analysis with errors-in-variables and a good alternative method to parametric approaches.

There is a strong connection between errors-in-variables and deconvolution problem, which has been studied extensively in the literature. See, for example, Carroll and Hall (1988), Fan (1991), Masry and Rice (1991), Stefanski and Carroll (1990) and Zhang (1990).

The paper is outlined as follows. Section 2 describes the deconvoluting kernel estimators and gives examples of these estimators according to the error distributions. Numerical examples based on continuous and binary responses with different error distributions are presented in Section 3. Section 4 discusses the limiting distributions of the proposed estimators and their consequences in providing confidence intervals. Section 5 contains some concluding remarks. The justifications of the limiting distribution of the deconvoluting kernel estimators are given in Section 6.

2 Methods

2.1 Kernel Estimators

Given a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y), where $X = X^\circ + \varepsilon$, we are interested in estimating the regression function of Y on the unobserved variable X°: $m(x) = E(Y|X^\circ = x)$. Note here the value of the covariate variable X_i° is not available, thus the estimating procedure must have the ability to extract X_i° from the observed $X_i = X_i^\circ + \varepsilon_i$.

Fan and Truong (1990) consider the following deconvoluting kernel estimator

$$\hat{m}_n(x) = \frac{\sum_j Y_j K_n(\frac{x - X_j}{h_n})}{\sum_i K_n(\frac{x - X_i}{h_n})} \tag{2.1}$$

with $K_n(x)$ given by

$$K_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx) \frac{\phi_K(t)}{\phi_\varepsilon(t/h_n)}\, dt. \tag{2.2}$$

Here $h_n \to 0$ is a smoothing parameter, and $\phi_K(\cdot)$ is the Fourier transform of the kernel function $K(\cdot)$:

$$\phi_K(t) = \int \exp(itx)K(x)\,dx,$$

and $\phi_\varepsilon(\cdot)$ is the characteristic function of the error variable ε.

Note that (2.1) is a kernel type estimator with kernel weights modified to account for the fact that covariates are measured with errors. The kernel function (2.2) can be motivated as follows. Suppose that we are interested in estimating the density function of the unobserved X°. Given a set of observations $X_1, \ldots, X_n$ with $X_i = X_i^\circ + \varepsilon_i$, Stefanski and Carroll (1990) and Fan (1991) consider the following deconvoluting density estimator:

$$\hat{f}_n(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\phi_K(th_n)\frac{\hat{\phi}_n(t)}{\phi_\varepsilon(t)}\,dt, \tag{2.3}$$

where $\hat{\phi}_n(\cdot)$ is the empirical characteristic function:

$$\hat{\phi}_n(t) = \frac{1}{n}\sum_1^n \exp(itX_j).$$

With notation (2.2), the estimator (2.3) can be written in the kernel form:

$$\hat{f}_n(x) = \frac{1}{nh_n}\sum_1^n K_n(\frac{x - X_j}{h_n}).$$

Now recall that the ordinary kernel regression estimator (i.e. no errors in covariates) is obtained from the kernel density estimator via the same line of reasoning. This leads to our deconvoluting estimator (2.1).

Throughout this paper, our discussion also based on the fact that the regression function has two derivatives.

Some theoretical aspects of the problem are investigated by Fan and Truong (1990), which can be highlighted as follows.

- The estimate (2.1) is optimal in terms of rate of convergence.
- The rates of convergence depend on the smoothness of error distributions, which can be characterized into two categories: ordinary smooth and super smooth. Let ϕ_ε be the characteristic function of the error distribution. Call a distribution

– super smooth of order β: if the characteristic function of the error distribution $\phi_\epsilon(\cdot)$ satisfies

$$d_0|t|^{\beta_0}\exp(-|t|^\beta/\gamma) \le |\phi_\epsilon(t)| \le d_1|t|^{\beta_1}\exp(-|t|^\beta/\gamma) \quad \text{as } t \to \infty, \tag{2.4}$$

where d_0, d_1, β, γ are positive constants and β_0, β_1 are constants;

– ordinary smooth of order β: if the characteristic function of the error distribution $\phi_\epsilon(\cdot)$ satisfies

$$d_0|t|^{-\beta} \le |\phi_\epsilon(t)| \le d_1|t|^{-\beta} \quad \text{as } t \to \infty, \tag{2.5}$$

for positive constants d_0, d_1, β.

If the second derivative of the regression function exists, the optimal rates of convergence is $O([\log n]^{-\frac{2}{\beta}})$ when the error distribution is super smooth of order β, and is $O(n^{-\frac{2}{2\beta+5}})$ when the error distribution is ordinary smooth of order β.

- The optimal choice of bandwidth h_n depends also on whether the error distribution is smooth or super smooth. For the super smooth error distribution of order β, it turns out that the optimal choice of $h_n = c(\log n)^{-1/\beta}$ for some constant c, which is known (see Example 2.1), and depends only on the error distribution and kernel function. In the ordinary smooth case, the optimal choice of bandwidth is $h_n = dn^{-\frac{1}{5+2\beta}}$, for some constant $d > 0$. The constant d is usually selected to balance the bias and variance of the estimate.

2.2 Examples

In this section, we give two examples of the deconvoluting kernel estimator (2.1) based on normal error and double exponential error distributions, which are super smooth of order 2 and ordinary smooth of order 2, respectively.

Example 2.1: Normal Errors

Suppose ε has a normal distribution with mean 0 and variance σ_0^2. Then $\phi_\varepsilon(t) = \exp(-\frac{1}{2}\sigma_0^2 t^2)$. Further, suppose the kernel $K(\cdot)$ is a function whose Fourier transform is given by

$$\phi_K(t) = (1-t^2)_+^3.$$

(This is a modified version of the inverse triangular kernel so that K is a second order kernel.) By (2.2),

$$K_n(x) = \frac{1}{\pi}\int_0^1 \cos(tx)(1-t^2)^3 \exp\left(\frac{\sigma_0^2 t^2}{2h_n^2}\right) dt. \tag{2.6}$$

Graphs of this function for different values of h_n are given Figure 1. According to Fan and Truong (1990), to achieve the optimal rates of convergence, the bandwidth h_n is chosen so that

$$h_n = c\sigma_0(\log n)^{-1/2} \quad \text{for } c \approx 1. \tag{2.7}$$

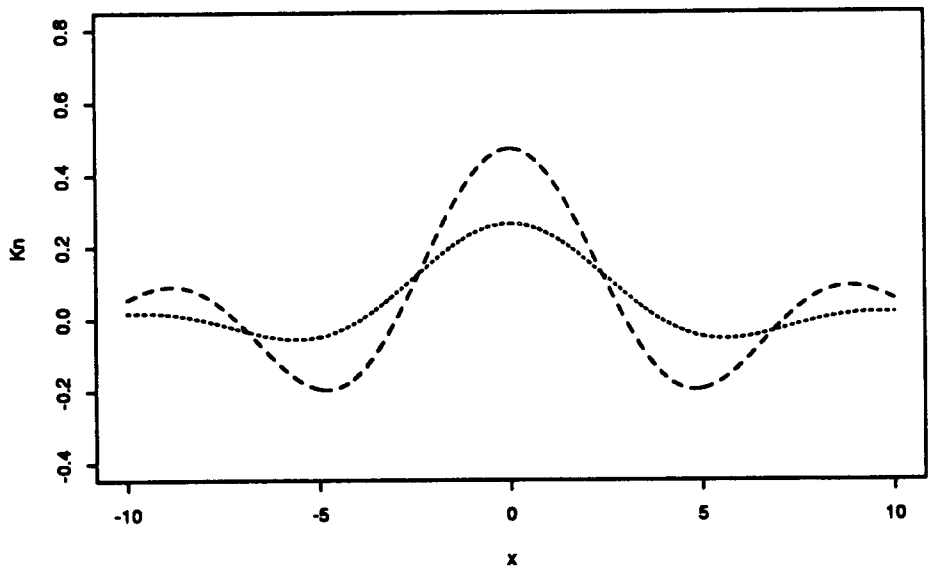

Figure 1. Deconvoluted kernel functions (2.6). *The dotted curve has* $c = 1.0$ *and the dash curve has* $c = 0.8$.

Example 2.2: Double Exponential Errors

Suppose ε has a double exponential distribution:

$$f_\varepsilon(z) = \sigma_0^{-1} \exp(-2|z|/\sigma_0). \tag{2.8}$$

Then

$$\phi_\varepsilon(t) = \frac{1}{1 + \frac{1}{4}\sigma_0^2 t^2}.$$

By (2.3),

$$\begin{aligned} K_n(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\phi_K(t)\left(1 + \frac{\sigma_0^2 t^2}{4h_n^2}\right) dt \\ &= K(x) + \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)t^2\phi_K(t)\, dt \\ &= K(x) - \frac{\sigma_0^2}{4h_n^2}K''(x). \end{aligned} \tag{2.9}$$

If $K(\cdot)$ is the Gaussian kernel $K(x) = (\sqrt{2\pi})^{-1}\exp(-x^2/2)$, then

$$K_n(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\left[1 - \frac{\sigma_0^2}{4h_n^2}(x^2 - 1)\right].$$

To obtain optimal rates of convergence, Fan and Truong (1990) showed that h_n is chosen so that

$$h_n \sim cn^{-1/9} \qquad \text{for some } c > 0.$$

2.3 Reliability Ratio

For the above examples, it is assumed that the variance of the error term σ_0^2 is known. This would seem rather restrictive at first. However, there are a number of situations, for example in psychology and sociology where information about σ_0^2 can be obtained via the battery of questions called *instrument* [see Fuller (1987, p.5)]. In fact, knowledge about σ_0^2 is usually expressed through the *reliability ratio*, which is defined by

$$r = \frac{\sigma_{X^\circ}^2}{\sigma_0^2 + \sigma_{X^\circ}^2} \approx 1 - \frac{\sigma_0^2}{s_n^2},$$

where s_n^2 is the sample variance for X, $\sigma_{X^\circ}^2$ is the variance of X°, and σ_0^2 is the variance of error distribution. See Fuller (1987) for a comprehensive account for regression analysis

based on the reliability ratio. Nevertheless, we are taking a different approach to this issue by viewing r as the signal-to-noise ratio in the sense that if the ratio is too small, the signal on covariate is too weak to give a meaningful analysis of the data. On the other hand, if the ratio is known to be about 90%, the above approximation gives $\sigma_0^2 = 0.1 s_n^2$, and hence the value of σ_0 is approximately known. In practice, one could analyze the data by trying, say $r = 0.5, 0.6, \ldots, 1.0$, and see which r would yield meaningful interpretations for the data being analyzed. $r \geq 0.5$ is chosen since we are trying to model the relationship between Y and X° rather than associating Y with the error term ε.

Comparing with parametric approaches, nonparametric approach is much more flexible. Indeed, for parametric approaches, in addition to specifying the reliability ratio, one further assumes a particular form of regression function, and a specific joint distribution between the covariate and the response. To judge the appropriateness of the model, one would have to follow the trial and error procedures, and this would lead to choosing parametric functions in an infinite dimensional space involving complicated computing in MLE. In contrast, our approach only requires to tune the signal to noise ratio r, which is much simpler. Another advantage is the shape of the deconvoluting kernel estimator (2.1) usually suggests a good parametric form for the regression function, in addition to being a useful diagnostic tool for regression analysis involving errors-in-variables.

2.4 Viewing Errors-in-Covariates as Errors-in-Responses

In this section, we give a new viewpoint of errors-in-variables problem, which provides some insight for understanding how errors in covariates would affect the scatterplot of the data. Suppose the regression function is smooth and the amount of error ε is small, or equivalently, $r = \sigma_{X^\circ}^2/(\sigma_0^2 + \sigma_{X^\circ}^2)$ is large. Then, it is clear that with high probability that the error ε contaminated in the true covariate $X^\circ (X = X^\circ + \varepsilon)$ is small. If X were the true design point, then its response at this point would be

$$Y^\circ = m(X) + \varepsilon^\circ$$

$$= \quad m(X^\circ + \varepsilon) + \varepsilon^\circ$$

$$\approx \quad m(X^\circ) + m'(X^\circ)\varepsilon + \varepsilon^\circ, \qquad \text{(Taylor's expansion)}$$

where ε° is the error in the regression model. But our observed response is $Y = m(X^\circ) + \varepsilon^\circ$. Thus, viewing covariate X as the true designed point, the observed response Y has an additional error $-m'(X^\circ)\varepsilon$, comparing with the true response Y° at point X.

There are two important consequences of the new view-point of the errors-in-variables:

- When the true regression curve is very flat (so that $|m'(\cdot)|$ is small), the errors-in-covariates do not appreciately affect the response. In other words, we can treat errors-in-covariates as if there were no error contaminated in the covariates. Thus, in linear regression model $m(x) = ax + b$, when $|a|$ is small, error in covariate does not affect greatly the ordinary linear regression analysis.

- When the reliability ratio is small, there is high probability that error ε contaminated in the covariate is small. Thus, viewing X as desired design point, the errors in the response are still negligible. In other words, error-in-covariate is negligible.

3 Simulation

3.1 Continuous Responses

In this section, we are interested in estimating the regression function $m(x) = x^3(1-x)^3$ [see, for example, Rice (1984)] based on a set of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ in which

$$Y_i = m(X_i^\circ) + \varepsilon_i^\circ, \quad \varepsilon_i^\circ \sim_{\text{iid}} N(0, 0.0015^2) \tag{3.1}$$

and the covariates X_i° are measured with errors according to

$$X_i = X_i^\circ + \varepsilon_i, \quad X_i^\circ \sim_{\text{iid}} \text{unif}(0,1). \tag{3.2}$$

Here $\{\varepsilon_i\}$ is a random sample from either $N(0, \sigma_0^2)$ or double exponential given by (2.8).

Before errors are added to the underlying covariates, it is helpful to show the scatterplot of 200 original observations (X_i^o, Y_i) (Figure 2) so that one can study the effect of measurement errors.

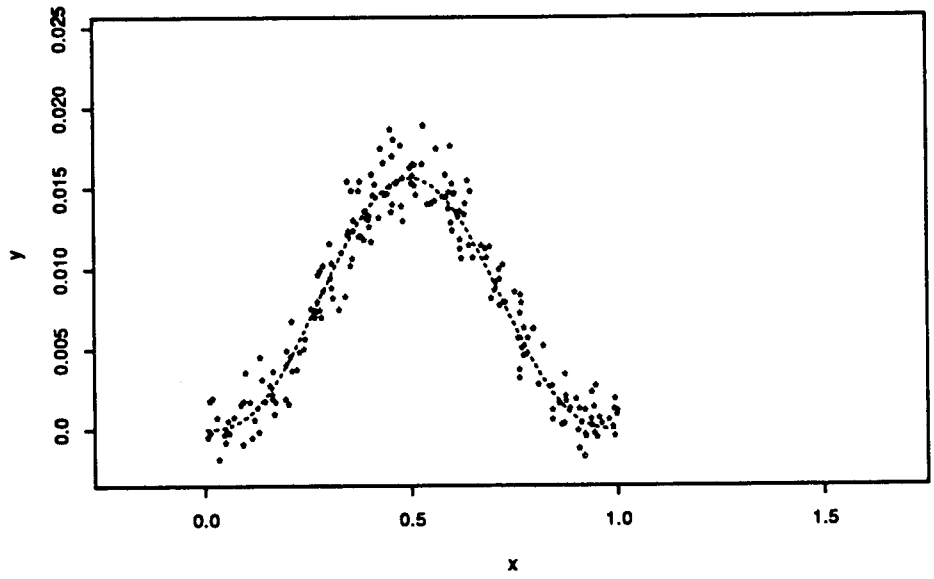

Figure 2. Scatterplot of the responses (3.1) *and the covariates without measurement errors.*

Example 3.1: Normal Errors

Figure 3 is the scatterplot of the responses Y_i and X_i by adding to the covariates X_i^o normal errors $N(0, \sigma_0^2)$ with σ_0^2 chosen so that the reliability ratio $r = 0.85$. Equivalently, these are normal errors with

$$\sigma_0^2 = \frac{1}{12} \cdot \frac{1-r}{r} = 0.0147.$$

The scatterplot looks noisier than Figure 2. This can easily be explained by viewing errors in covariates as additional errors in the responses. See Section 2.4. The regression function is then estimated by the deconvoluting kernel estimator (2.1) with different degrees of smoothing by choosing the following bandwidths: (see (2.7))

$$h_n = \sigma_0(\log n)^{-1/2}, \qquad h_n = 0.8\sigma_0(\log n)^{-1/2}.$$

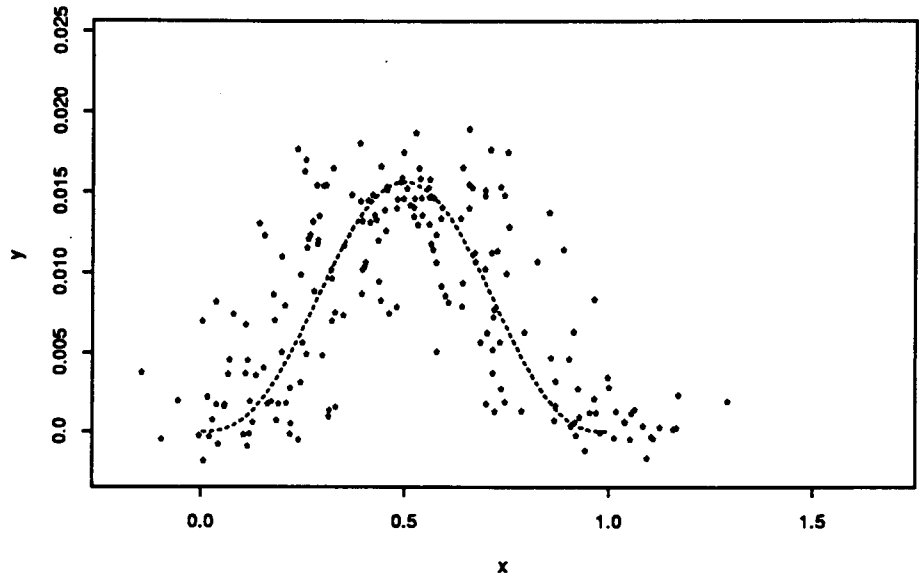

Figure 3. Scatterplot of the responses and the covariates contaminated with normal errors ($r = 0.85$).

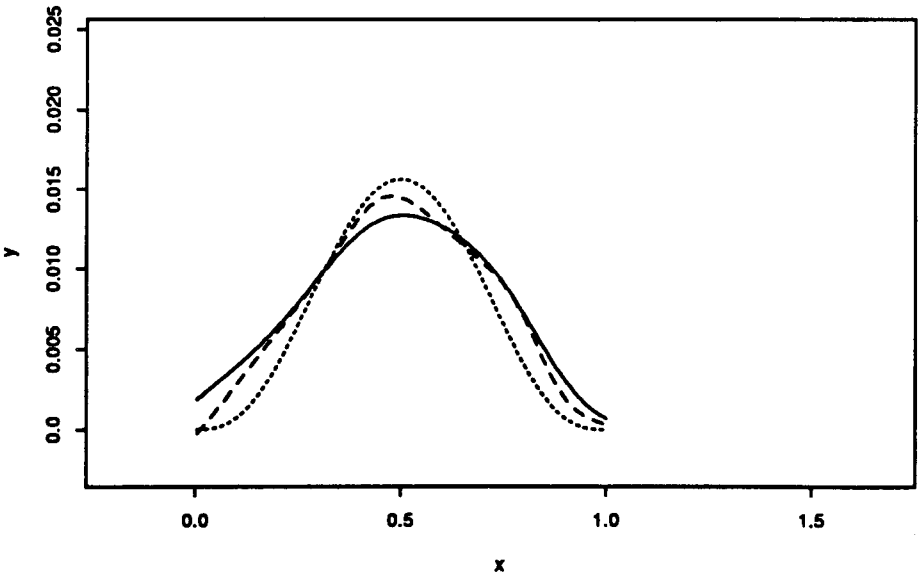

Figure 4. Deconvoluted kernel estimates: normal errors with $r = 0.85$. *The solid and the dash curves are the estimates with* $c = 1.0$ *and* $c = 0.80$, *respectively. The estimates are evaluated at* $x = 1/200, 2/200, \ldots, 1$. *The dotted curve is the true regression function.*

Figure 4 shows the estimates (at $x = 1/200, 2/200, \ldots, 1$) superimposed with the underlying regression function $m(x) = x^3(1-x)^3$. Note that in this case, the bandwidth $h_n = \sigma_0(\log n)^{-1/2}$ provides a threshold for nonparametric estimation involving normal errors. If $h_n = c\sigma_0(\log n)^{-1/2}$ is chosen with $c > 1$, then the variance converges to 0 much faster than the bias term. If c would have been chosen to be much less than 1, then the variance can be quite large. In fact, it can even diverge to ∞.

The graphs of the corresponding deconvoluting kernel functions used in the above estimates are indicated in Figure 1. Note that in contrast to the ordinary approach, these kernel functions depend on n.

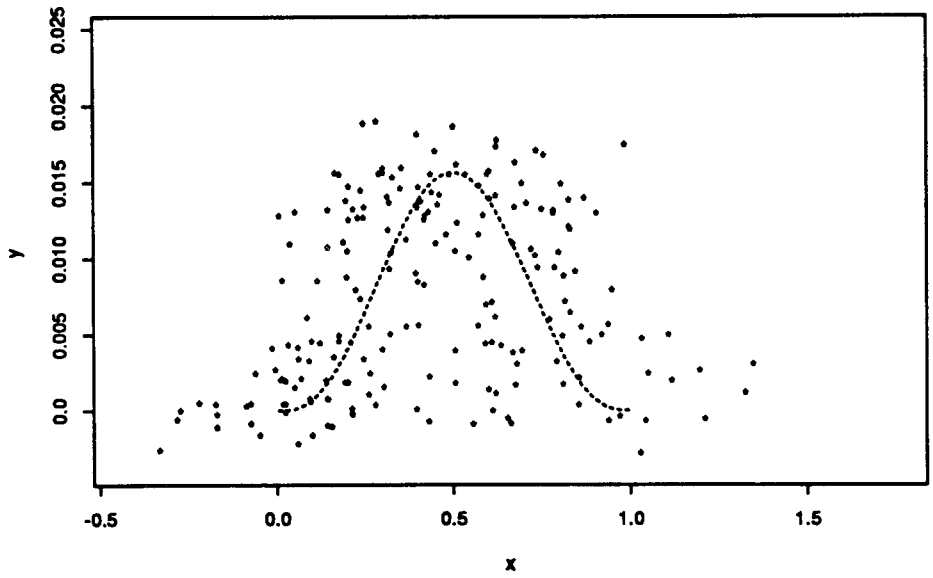

Figure 5. Scatterplot of the responses and the covariates contaminated with normal errors ($r = 0.70$).

Figures 5–6 show the results when the variance of the error distribution is larger by decreasing the reliability ratio r to 0.70. The resulting scatterplot is much noisier than the one in Figure 3 with $r = 0.85$. The effect of choosing $c < 1$ (hence inflating the variance) is more transparent, see the estimates with $c = 0.6$ and $c = 0.8$ in Figure 6. This is compatible

with the theory in Fan and Truong (1990).

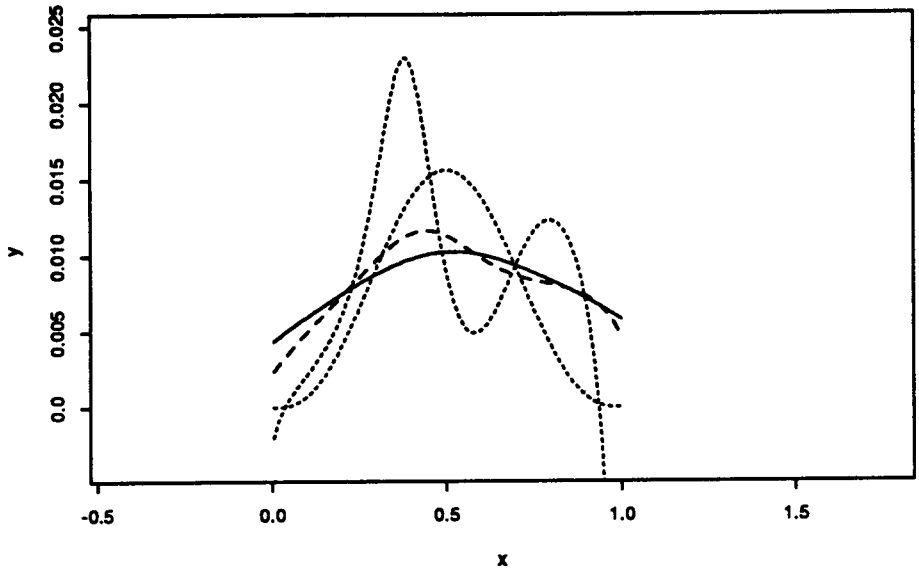

Figure 6. Deconvoluted kernel estimates: normal errors with $r = 0.70$. The solid, dash and the dotted curves are the estimates with $c = 1.0$, $c = 0.80$ and $c = 0.60$, respectively. The true regression function is also plotted for comparison.

Example 3.2: Double Exponential Errors

Figure 7 shows the data with double exponential errors added to the covariates. These errors are generated from (2.8) with σ_0 chosen so that $r = 0.80$.

The regression function is then estimated by smoothing the data in Figure 7 using the deconvoluting kernel (2.9). The results are presented in Figure 8.

Figures 9–10 illustrate the effect of having a larger variance (since $r = 0.60$) in the double exponential error distribution. In contrast with normal errors, even though the reliability ratios are smaller, the resulting estimates are closer to the true curve: regression with double exponential errors is easier than the regression with normal errors in the covariates. This is consistent with the remarks in Section 2.

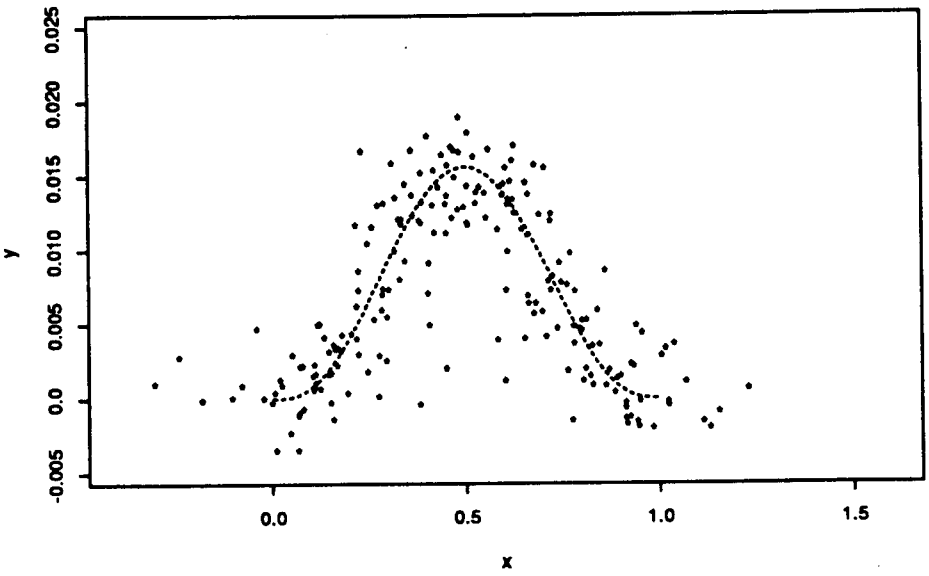

Figure 7. Scatterplot of the responses and the covariates contaminated with double exponential errors ($r = 0.80$).

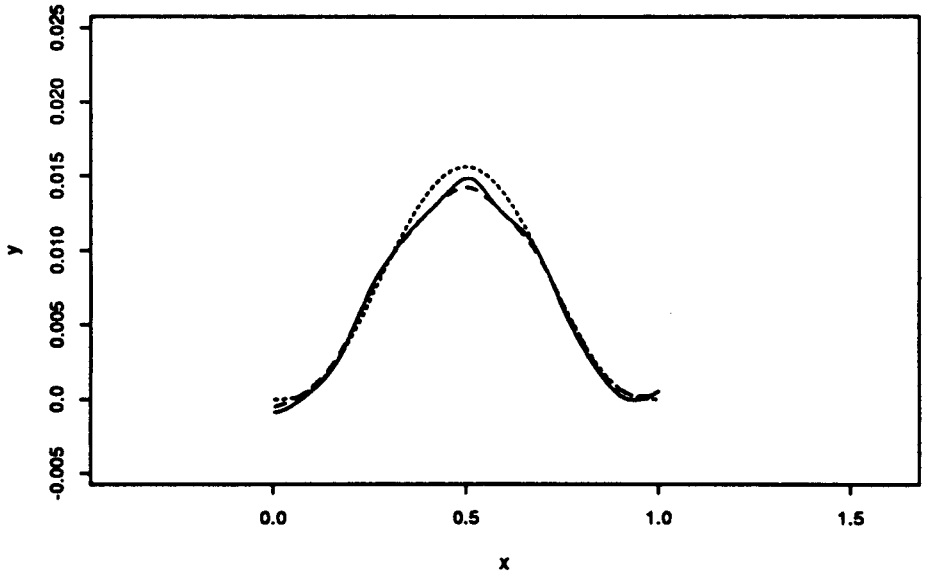

Figure 8. Deconvoluted kernel estimates: double exponential errors with $r = 0.80$. *The solid and dash curves are the estimates with* $h_n = 0.07$ *and* $h_n = 0.08$, *respectively. The dotted curve is the true regression function.*

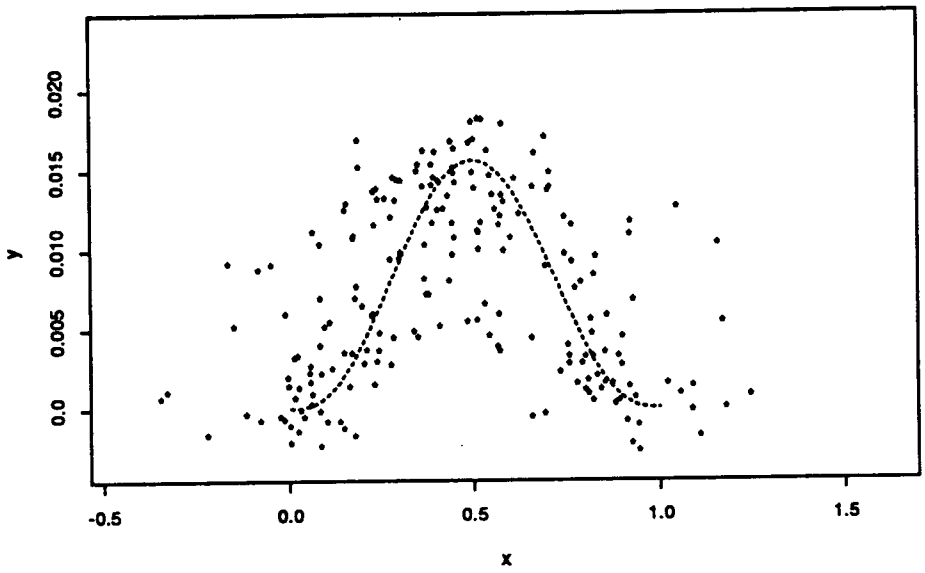

Figure 9. Scatterplot of the responses and the covariates contaminated with double exponential errors ($r = 0.60$).

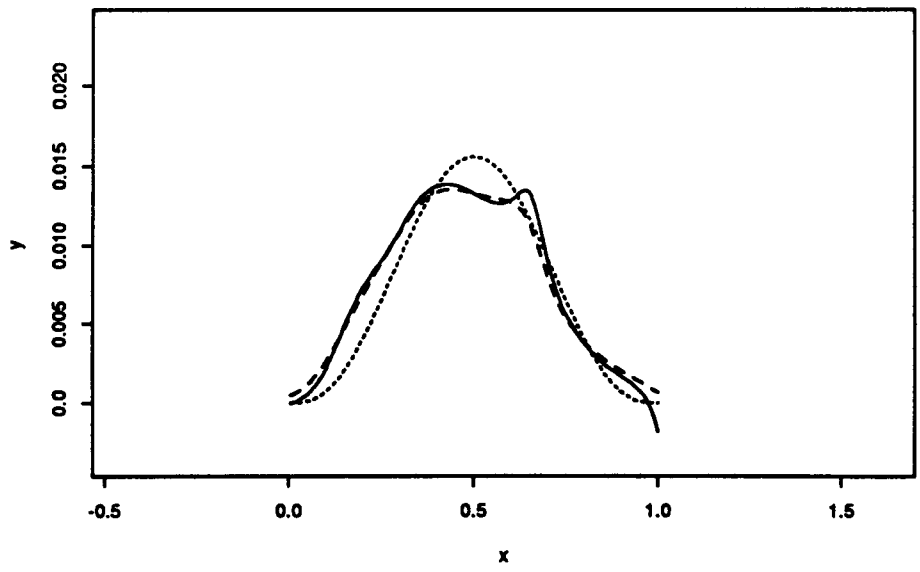

Figure 10. Deconvoluted kernel estimates: double exponential errors with $r = 0.60$. *The solid and dash curves are the estimates with* $h_n = 0.09$ *and* $h_n = 0.10$, *respectively. The dotted curve is the underlying regression function* $m(\cdot)$.

3.2 Binary Responses

Many interesting applications in medical and epidemiologic studies involve binary responses. See, for example, Carroll et al (1984), Prentice (1986) and Whittemore and Keller (1988). The nonparametric estimators considered in the previous sections can also be applied to this type of response as well.

Let Y denote a 0-1 random variable with distribution depending on the covariate $X^\circ = x$ so that the logit of Y on $X^\circ = x$ is given by

$$\text{logit}(Y|X^\circ = x) = 8(x - 0.5)^3.$$

That is, the regression function is given by

$$m(x) = E(Y|X^\circ = x) = P(Y = 1|X^\circ = x) = \frac{\exp(8(x-0.5)^3)}{1 + \exp(8(x-0.5)^3)}.$$

As before, let X° have a uniform distribution on (0,1). X° is not available and it is observed through $X = X^\circ + \varepsilon$. The distribution of the error ε is treated separately as follows.

Example 3.3: Double Exponential Errors

Let $(X_1, Y_1), \ldots, (X_{200}, Y_{200})$ denotes a random sample from the distribution of (X, Y) with ε having a double exponential distribution (2.8). Here σ_0 is chosen so that the reliability ratio $r = 0.90$. Thus

$$\begin{aligned} X_i &= X_i^\circ + \varepsilon_i, \quad X_i^\circ \sim_{\text{iid}} \text{unif}(0,1), \quad \varepsilon_i \sim_{\text{iid}} f_\varepsilon, \quad \text{[see (2.8)]} \\ Y_i &= \begin{cases} 1, & \text{with probability } m(X_i^\circ); \\ 0, & \text{with probability } 1 - m(X_i^\circ). \end{cases} \end{aligned}$$

The data are presented in Figure 11, there the estimates (at $x = 1/200, 2/200, \ldots, 199/200, 1$) are obtained via (2.1) with the deconvoluting kernel function given by (2.9) and the bandwidth $h_n = 0.10$.

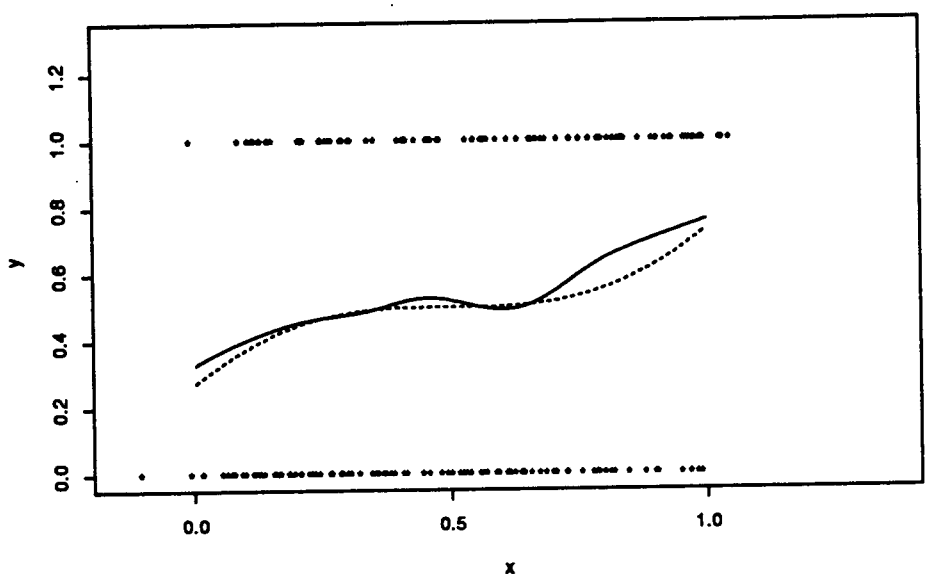

Figure 11. Scatterplot of the responses and the covariates contaminated with double exponential errors ($r = 0.90$). *The solid curve is the estimate* (2.1) *and* (2.9) *with* $h_n = 0.1$. *The dotted curve is the underlying regression function* $\text{logit}(Y|X^\circ = x) = 8(x - 0.5)^3$.

Example 3.4: Normal Errors

In this example, let X° denote a unif$(0,1)$ random variable and ε has a normal distribution $N(0,\sigma_0^2)$ with σ_0^2 chosen so that the reliability ratio $r = 0.90$. $(X_1,Y_1),\ldots,(X_{200},Y_{200})$ is a random sample from the distribution of (X,Y), where $X = X^\circ + \varepsilon$. The scatterplot of the data is shown in Figure 12. Choose h_n by setting $c = 1.3$ in (2.7), we obtain the deconvoluting kernel estimates (at $x = 1/200, 2/200, \ldots, 199/200, 1$) via (2.1) and (2.6). See Figure 12.

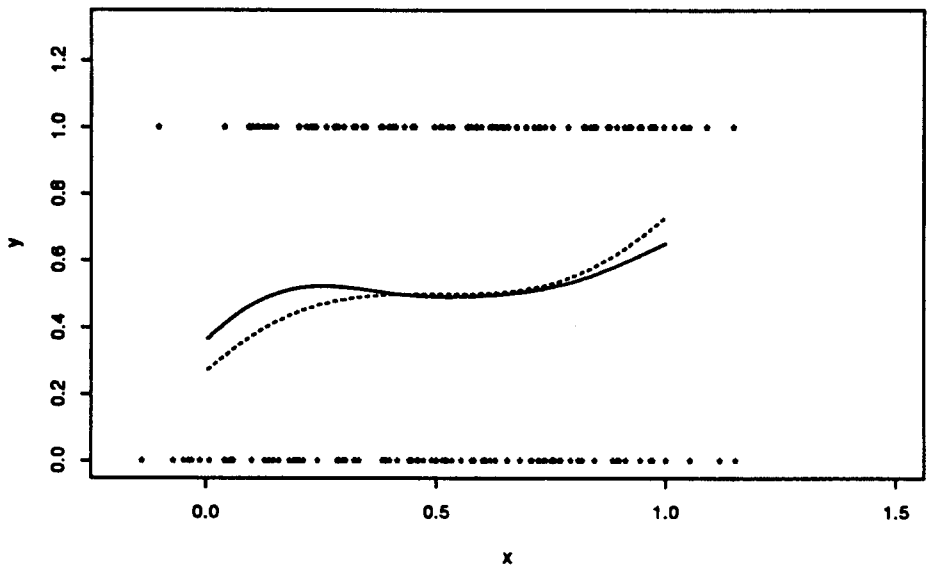

Figure 12. Scatterplot of the responses and the covariates contaminated with normal errors ($r = 0.90$). *The solid curve is the estimate* (2.1) *and* (2.6) *with* $h_n = 0.06$. *The dotted curve is the underlying regression function* $\text{logit}(Y|X^\circ = x) = 8(x - 0.5)^3$.

4 Confidence Intervals

To construct confidence intervals for the unknown regression function in the presence of errors in variables, we need to validate the asymptotic normality. For simplicity of notation, denote

$$K_{nj}(x) = \frac{1}{h_n} K_n \left(\frac{x - X_j}{h_n} \right), \tag{4.1}$$

which is the deconvolution weight of estimator (2.1). Let

$$b_n(x) = \frac{1}{h_n} \int_{-\infty}^{\infty} [m(u) - m(x)] K(\frac{x-u}{h_n}) f_X(u)\, du, \tag{4.2}$$

where $f_X(\cdot)$ is the density of X_j. For the bandwidth and the kernel that we will use, we have

$$\hat{f}_n(x) \xrightarrow{P} f_{X^\circ}(x), \tag{4.3}$$

where $\hat{f}_n(x)$ is defined by (2.3). Thus,

$$\begin{aligned}\hat{m}_n(x) - m(x) &= \frac{1}{n}\sum_1^n (Y_j - m(x))K_{nj}(x)/f_{X^\circ}(x)(1 + \frac{f_{X^\circ}(x)}{\hat{f}_n(x)} - 1) \\ &= \frac{1}{n}\sum_1^n (Y_j - m(x))K_{nj}(x)/f_{X^\circ}(x)(1 + o_P(1)). \qquad (4.4)\end{aligned}$$

Hence the asymptotic normality for $\hat{m}_n(x) - m(x)$ is equivalent to the asymptotic normality for $n^{-1}\sum_1^n (Y_j - m(x))K_{nj}(x)$, which is the mean of i.i.d. random variables.

We begin with some assumptions.

CONDITIONS

1. The marginal density $f_{X^\circ}(\cdot)$ of the unobserved X° is bounded away from zero at point x, and has a continuous bounded derivative.

2. The regression function $m(\cdot)$ has a bounded second derivative.

3. The conditional variance of $\sigma^2(x) = \mathrm{var}(Y|X^\circ = x)$ is bounded away from zero and infinity. Moreover, there is a $\delta > 0$ such that $\mu_{2+\delta}(x) \equiv E\left(|Y|^{2+\delta}|X^\circ = x\right) < \infty$.

4. The characteristic function $\phi_\varepsilon(\cdot)$ of the error variable ε does not vanish.

Let's first consider double exponential errors. Assume more generally that there exist constants $c \neq 0$, and $\beta \geq 0$ such that

$$\phi_\varepsilon(t)t^\beta \to c, \quad \phi'_\varepsilon(t)t^{\beta+1} \to -\beta c. \qquad (4.5)$$

Then, it is very easy to verify that the double exponential distribution (2.8) is a special case of (4.5).

Call a kernel K a second order kernel, if K is bounded and satisfies

$$\int_{-\infty}^{\infty} K(x)dx = 1, \int_{-\infty}^{\infty} xK(x)dx = 0, \int_{-\infty}^{\infty} x^2K(x) \neq 0.$$

Moreover, we always assume that

$$\int_{-\infty}^{\infty} [|\phi_K(t)| + |\phi'_K(t)|]|t|^{2\beta} < \infty,$$

where $\phi_K(\cdot)$ is the Fourier transform of the kernel K.

Now, we are ready to study the asymptotic normality for double exponential error distributions.

Theorem 1. *Under Conditions* 1–4 *and* (4.5), *if the kernel function* $K(\cdot)$ *is a second order kernel, then*

$$\sqrt{n}\frac{f_{X^\circ}(x)[\hat{m}_n(x) - m(x)] - b_n(x)}{\sqrt{E\sigma^2(X_1)K_{n1}^2(x)}} \longrightarrow N(0,1), \tag{4.6}$$

provided that $h_n \to 0$ *and* $nh_n^{1+2\beta} \to \infty$.

A practical version of Theorem 1 involves how to estimate bias and variance in (4.6). This is usually done as follows by letting h_n converges 0 a little bit faster than the optimal choice [which is $h_n = O(n^{-\frac{1}{2\beta+5}})$ by Fan and Truong (1990)] so that the bias is negligible. To avoid technical difficulties, assume $\sigma^2(x) = \sigma^2$, which is independent of x. Then, the asymptotic variance is given by

$$E\sigma^2(X_1)K_{n1}^2(x) = \sigma^2 EK_{n1}^2(x).$$

Let's take the interval $[a,b]$ such that the marginal density f_{X° satisfies $\min_{x\in[a,b]} f_{X^\circ}(x) > 0$. Then σ^2 can be estimated by

$$\hat{\sigma}_n^2 = \sum_1^n (Y_j - \hat{m}(X_j))^2 1_{[a\le X_j\le b]} / \sum_1^n 1_{[a\le X_j\le b]}$$

and the second factor can be estimated by $\frac{1}{n}\sum_1^n K_{nj}^2(x)$, see (4.1. Denote

$$\hat{V}_n^2(x) = \hat{\sigma}_n^2 \frac{1}{n}\sum_1^n K_{nj}^2(x). \tag{4.7}$$

Corollary 1. *Under the conditions of Theorem* 1 *and the assumption that* $\sigma^2(x) = \sigma^2$, *if* $h_n = o(n^{-\frac{1}{2\beta+5}})$, *and* $nh_n^{1+2\beta} \to \infty$, *then*

$$\sqrt{n}\frac{\hat{f}_n(x)[\hat{m}_n(x) - m(x)]}{\hat{V}_n(x)} \longrightarrow N(0,1). \tag{4.8}$$

As a direct consequence, the $(1-\alpha)$ confidence interval for $m(x)$ is given by

$$m(x) \in \hat{m}_n(x) \pm z_{1-\alpha/2} \frac{\hat{V}_n(x)}{\sqrt{n}\hat{f}_n(x)}, \tag{4.9}$$

where $z_{1-\alpha}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

Example 3.2: Double Exponential Errors (cont'd)

Recall that in this case, the covariates are measured with double exponential errors ($r = 0.80$). Based on the estimate (2.1) and (2.9) with $h_n = 0.08$ (Figure 8), the 95% pointwise confidence intervals (4.9) are given in Figure 13. It is interesting to note that the whole regression curve $m(x) = x^3(1-x)^3$ lies within the confidence limits.

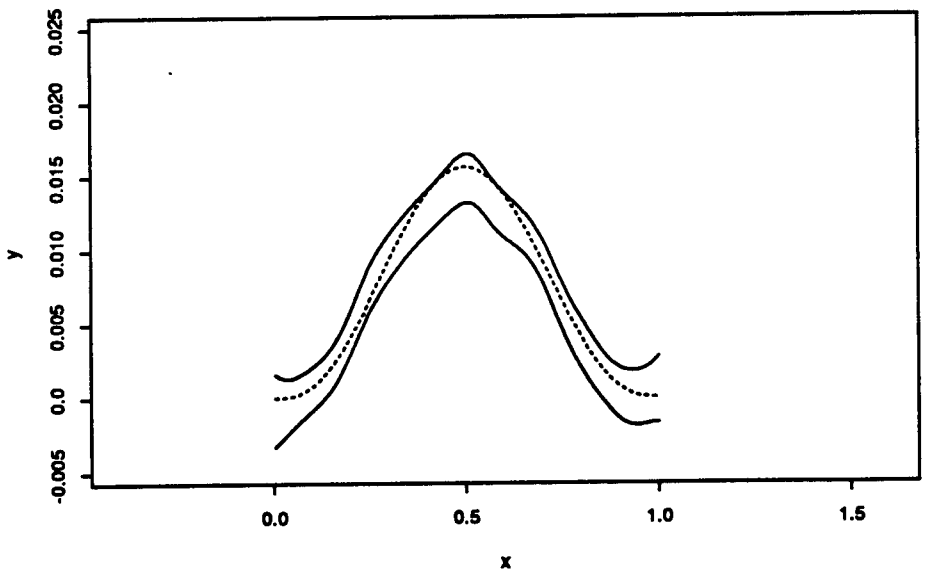

Figure 13. 95% pointwise confidence intervals for $m(x) = x^3(1-x)^3$. Based on the data in Figure 7.

Now, let's state the asymptotic normality for normal error distributions.

Theorem 2. *Under Conditions* 1–4, *if* $\varepsilon \sim N(0, \sigma_0^2)$, *then for the kernel estimator* (2.1) *with* K_n *given by* (2.6), *we have*

$$\sqrt{n}\frac{f_{X^*}(x)[\hat{m}_n(x) - m(x)] - b_n(x)}{\sqrt{E\sigma^2(X_1)K_{n1}^2(x)}} \longrightarrow N(0,1), \tag{4.10}$$

provided that $h_n = d\sigma_0(\log n)^{-1/2}$ *with* $d \geq 1 - \frac{3\log\log n}{\log n}$.

Note that the condition $d \geq 1 - \frac{3\log\log n}{\log n}$ is used to guarantee that (4.3) holds. To wit, note that for K_n defined by (2.6),

$$\begin{aligned} |K_n(\cdot)| &\leq \int_0^1 (1-t^2)^3 \exp\left(\frac{\sigma_0^2 t^2}{2h_n^2}\right) dt \\ &\leq \exp\left(\frac{(1-h_n^{1.8})^2\sigma_0^2}{2h_n^2}\right) + h_n^{5.4}\int_{1-h_n^{1.8}}^1 2t\exp\left(\frac{\sigma_0^2 t^2}{2h_n^2}\right) dt \\ &= \frac{2h_n^{7.4}}{\sigma_0^2}\exp\left(\frac{\sigma_0^2}{2h_n^2}\right)(1+o(1)). \end{aligned} \tag{4.11}$$

Thus,

$$\begin{aligned} \mathrm{var}(\hat{f}_n(x)) &\leq \frac{1}{n}E|K_{n1}|^2 \\ &= O\left(\frac{1}{n}h_n^{12.8}\exp\left(\frac{\sigma_0^2}{h_n^2}\right)\right) \\ &= O\left(\frac{1}{n(\log n)^{6.4}}\exp\left(\log n\left[1 - \frac{3\log\log n}{\log n}\right]^{-2}\right)\right) \\ &\to 0. \end{aligned}$$

Now the bias has the expression (see Fan(1991))

$$E\hat{f}_n(x) - f_{X^*}(x) = \int_{-\infty}^{\infty}[f_{X^*}(x - h_n y) - f_{X^*}(x)]\,K(y)dy \to 0.$$

Thus, (4.3) follows since the bias and the variance of $\hat{f}_n(x)$ both converge to zero.

Corollary 2. *Under the conditions of Theorem* 2, *if* $\sigma^2(x) = \sigma^2$ *and*

$$h_n = \left(1 - \frac{3\log\log n}{\log n} + o\left(\frac{3\log\log n}{\log n}\right)\right)\sigma_0(\log n)^{-1/2},$$

we have

$$\sqrt{n}\frac{\hat{f}_n(x)[\hat{m}_n(x) - m(x)]}{\hat{V}_n(x)} \longrightarrow N(0,1), \tag{4.12}$$

where $\hat{V}_n(x)$ *is defined by* (4.7).

For normal error distribution, the confidence interval for $m(x)$ has a similar form as (4.9).

Example 3.1: Normal Errors (cont'd)

Here the covariates are measured with normal errors ($r = 0.85$). Using (4.9) and the estimate (2.1), (2.6), (2.7) with $c = 0.8$ (see Figure 4), 95% pointwise confidence intervals are presented in Figure 14. Note that the regression curve does not lie totally within the confidence limits (compare with double exponential error case), this may be associated with the fact that the rate of convergence of the deconvoluting kernel estimators (2.1) is very slow. See Fan and Truong (1990).

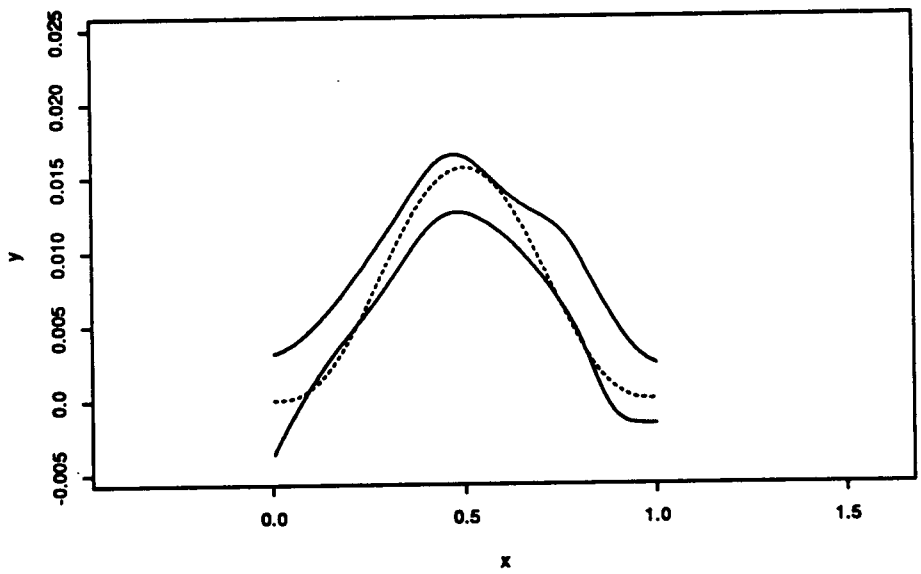

Figure 14. 95% pointwise confidence intervals for $m(x) = x^3(1-x)^3$. *Based on the data in Figure 3.*

Note that for the normal error distributions, in order to have the variance of the deconvolution kernel estimator converges to zero, the bandwidth must satisfy $h_n = d\sigma_0(\log n)^{-1/2}$

with $d \geq 1$, while for the variance to have lower order than the bias, one requires that $h_n = d\sigma_0(\log n)^{-1/2}$ with $d \leq 1$. This determines the bandwidth in Corollary 2.

For the binary response models in Section 3.2, $m(x)$ is the conditional probability $P(Y = 1|X^\circ = x)$ with conditional variance given by $m(x)(1-m(x))$, which depends on x. Hence, the confidence intervals (4.9) should be modefied to account for the heteroscedasticity as follows:

$$m(x) \in \hat{m}_n(x) \pm z_{1-\alpha/2} \frac{\sqrt{\sum K_{nj}^2(x)}\sqrt{\{\hat{m}_n(x)(1-\hat{m}_n(x))\}_+}}{n\hat{f}_n(x)}.$$

Figures 15–16 show the 95% confidence intervals based on the binary response data presented in Examples 3.3 and 3.4. See also Figures 11 and 12.

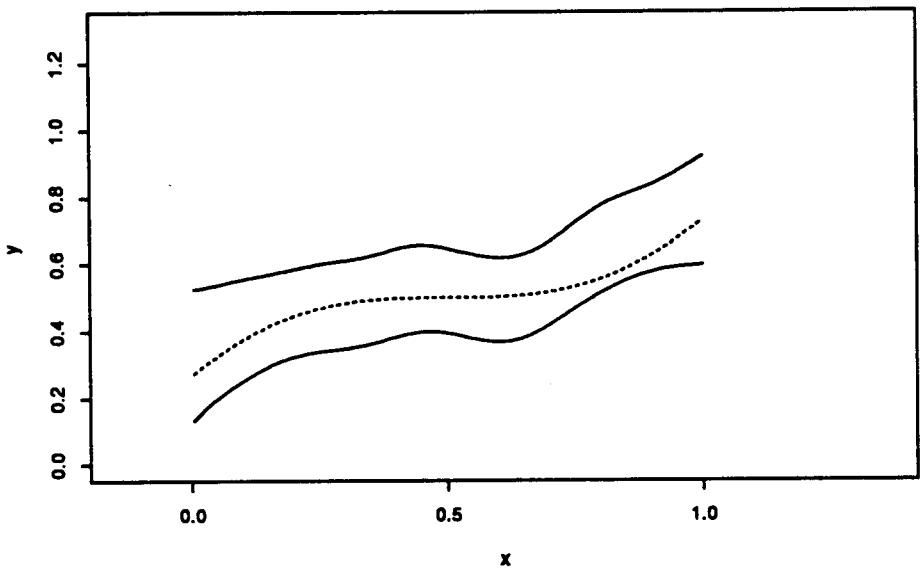

Figure 15. 95% pointwise confidence intervals for $\text{logit}(Y|X^\circ = x) = 8(x - 0.5)^3$. *Based on the binary response data in Figure 11.*

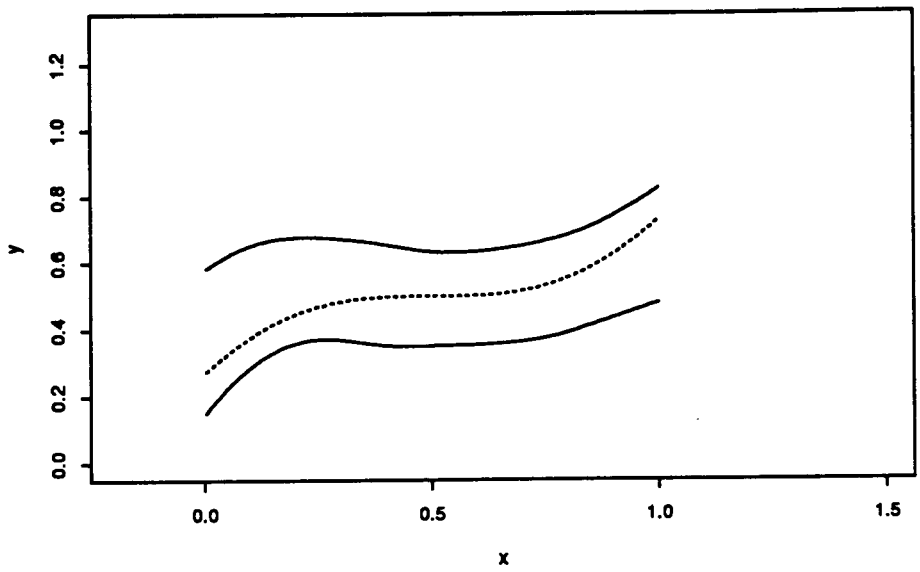

Figure 16. 95% pointwise confidence intervals for $\text{logit}(Y|X^\circ = x) = 8(x - 0.5)^3$. *Based on the binary response data in Figure 12.*

5 Conclusions

Much work has been done on the errors-in-variables problem under parametric assumptions for the regression function, and the error distribution is usually taken to be Gaussian. Although recent attention has turned to other error distributions and nonlinear (parametric) regression model (Carroll et al. (1984) and Prentice (1986)), the computation needed, however, to obtain maximum likelihood estimates are arduous and intractable [Whittemore (1989)]. This difficulty has prompted to the development of method based on approximation. Still, there are further problems along this line. Namely, the sampling properties of these approximating procedures are not easy to obtain (Stefanski (1985), Whittemore and Keller (1988)) and there are no diagnostic tools. Such important methodology is needed in, for instance, Examples 3.3 and 3.4 on nonlinear logistic regression analysis.

In contrast, we proposed nonparametric estimators based on the idea of deconvolution. These estimates are *easy to compute*, *flexible* in describing the relationship for the response and the covariate, they are also useful *diagnostic tools* for data analysis involving errors-in-variables. More importantly, both the sampling properties of these estimators and confidence intervals are established. The analyses given in the above examples have demonstrated the importance of these features, which also make the nonparametric procedures potentially useful for a board class of problems.

Further work in this direction may be needed to address the issue on bandwidth selection. Our simulations showed that the visual bandwidth selection (i.e. bandwidths chosen by the users) performed reasonably well. Methods based on cross-validation appear promising and offer interesting research problems.

6 Appendix—proofs

6.1 Proofs of Theorem 1 & 2

Denote [see (4.4)]

$$U_{nj} = (Y_j - m(x))K_{nj}(x). \tag{6.1}$$

Note that (4.4) is the sum of i.i.d. random variables. By the triangular central limit theorem,

$$\sqrt{n}\frac{\frac{1}{n}\sum_1^n U_{nj} - EU_{n1}}{\sqrt{\text{var}(U_{n1})}} \longrightarrow N(0,1), \tag{6.2}$$

if Lyapunov's condition holds:

$$\frac{E\,|U_{n1} - EU_{n1}|^{2+\delta}}{n^{\delta/2}[\text{var}(U_{n1})]^{1+\delta/2}} \to 0. \tag{6.3}$$

It has been calculated (see Lemma 6.2 of Fan and Truong (1990)) that

$$EU_{n1} = \frac{1}{h_n}\int_{-\infty}^{\infty}[m(u) - m(x)]K(\frac{x-u}{h_n})f_X(u)\,du = b_n(x). \tag{6.4}$$

A direct consequence of (6.4) is that the sequence $\{EU_{n1}\}$ is bounded. Since the sequence $\{E|U_{n1}|^\alpha\}(\alpha = 2, 2+\delta)$ is an unbounded sequence, the condition (6.3) is equivalent to

$$\frac{E\,|U_{n1}|^{2+\delta}}{n^{\delta/2}[EU_{n1}^2]^{1+\delta/2}} = \frac{E\,|Y_1 - m(x)|^{2+\delta}\,|K_{n1}(x)|^{2+\delta}}{n^{\delta/2}[E\,|(Y_1 - m(x))K_{n1}(x)|^2]^{1+\delta/2}} \to 0, \tag{6.5}$$

where $K_{n1}(x)$ is defined by (4.1). By Condition 3, the right hand side of (6.5) is of order

$$\frac{E\mu_{2+\delta}(X_1)|K_{n1}(x)|^{2+\delta}}{n^{\delta/2}\,[E\sigma^2(X_1)K_{n1}^2(x)]^{1+\delta/2}} \le C\frac{E|K_{n1}(x)|^{2+\delta}}{n^{\delta/2}\,[EK_{n1}^2(x)]^{1+\delta/2}}, \tag{6.6}$$

for some constant C. If we use the same argument for asymptotic normality of the deconvolution problem (estimating the density $f_{X^\bullet}$), we end up with exactly the same problem: showing that (6.6) converges to zero, which has been justified by Fan (1990). Thus, Lyapounov's condition (6.3) holds. Since

$$EU_{n1} = o(EU_{n1}^2), \quad \mathrm{var}(U_{n1}) = EU_{n1}^2(1+o(1)) = E\sigma^2(X_1)K_{n1}^2(x)(1+o(1)).$$

Thus, by (6.2), (6.4) and (4.4), the conclusion of Theorem 1 & 2 follows.

6.2 Proof of Corollary 1

With bandwidth selection in Corollary 1, the bias term is negligible [see Fan and Truong (1990)]:

$$\frac{\sqrt{n}b_n(x)}{\sqrt{E\sigma^2(X_1)K_{n1}^2(x)}} = O\left(\sqrt{n}h_n^2/\sqrt{h_n^{-2\beta-1}}\right) = o(1). \tag{6.7}$$

Consequently,

$$\frac{\sqrt{n}f_{X^\bullet}(x)(\hat{m}_n(x) - m(x))}{\sqrt{\sigma^2 EK_{n1}^2(x)}} \to N(0,1).$$

Thus, we need only to show

$$\frac{\hat{V}_n^2}{\sigma^2 EK_{n1}^2(x)} \xrightarrow{P} 1.$$

According to the Weak Law of Large Number [Chow and Teicher (1978), p. 328]

$$\frac{\frac{1}{n}\sum_1^n K_{nj}^2(x)}{EK_{n1}^2(x)} \xrightarrow{P} 1$$

holds, if for each $\varepsilon > 0$,

$$\frac{1}{EK_{n1}^2(x)} E\left(K_{n1}^2(x) 1_{[K_{n1}^2(x) \geq \varepsilon n E K_{n1}^2(x)]}\right) \to 0, \tag{6.8}$$

which is proved by Fan (1990) in the deconvolution setting. Thus, to complete the proof, we need only to show that $\sigma_n^2 \xrightarrow{P} \sigma^2$. Since the density $f_{X^*}(\cdot)$ is bounded away from 0 on $[a, b]$, we have

$$\sup_{a \leq x \leq b} |\hat{m}_n(x) - m(x)| \xrightarrow{P} 0. \tag{6.9}$$

Thus, by (6.9), we have

$$\begin{aligned}
& \frac{1}{n}\sum_1^n (Y_j - \hat{m}_n(X_j))^2 1_{[a \leq X_j \leq b]} \\
= \; & \frac{1}{n}\sum_1^n (Y_j - m(X_j))^2 1_{[a \leq X_j \leq b]} + o_P(1) \\
= \; & E(Y_1 - m(X_1))^2 1_{[a \leq X_j \leq b]} + o_P(1) \\
= \; & \sigma^2 P(a \leq X_1 \leq b) + o_P(1).
\end{aligned}$$

and

$$\frac{1}{n}\sum_1^n 1_{[a \leq X_j \leq b]} \xrightarrow{P} P(a \leq X_1 \leq b).$$

Thus,

$$\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2.$$

The conclusion follows.

6.3 Proof of Corollary 2

Similarly to the proof of Corollary 1, we need only to show that both (6.7) and (6.8) hold for K_n given by (2.6). To this end, for $x \in [0, \pi/4]$,

$$\begin{aligned}
K_n(x) &\geq \frac{1}{2\pi}\int_{1-h_n^2}^1 (1-t^2)^3 \exp\left(\frac{\sigma_0^2 t^2}{2h_n^2}\right) dt \\
&\geq \frac{1}{2\pi}\exp\left(\frac{\sigma_0^2(1-h_n^2)^2}{2h_n^2}\right)\int_{1-h_n^2}^1 (1-t)^3 dt \\
&\geq \frac{1}{8\pi} h_n^8 \exp\left(\frac{\sigma_0^2}{2h_n^2} - \sigma_0^2\right).
\end{aligned} \tag{6.10}$$

Let $C = \frac{1}{8\pi}\exp(-\sigma_0^2)$. Note that the density of X is the convolution of f_{X^*} with a normal distribution. Thus, $P(X \in [0, \pi/4]) > 0$. By (6.10),

$$EK_{n1}^2(x) \geq C^2 h_n^{14} \exp(\sigma_0^2/h_n^2) P(X \in [0, \frac{\pi}{4}]). \tag{6.11}$$

It follows from (6.11) that (6.7) holds. Using (6.11) together with (4.11), the set $\{K_{n1}^2(x) \geq \varepsilon n E K_{n1}^2(x)\}$ is an empty set, when n is large. Hence (6.8) holds.

References

[1] Armstrong, B. (1985). Measurement error in the generalized linear models. *Communications in Statistics — Simulation and Computation*, 14, 529–544.

[2] Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 70, 19–25.

[3] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvoluting a density. *J. Amer. Statist. Assoc.*, 83, 1184–1186.

[4] Chow, Y. S. and Teicher, H. (1978). *Probability Theory: independence, interchangeability, martingales.* Springer Verlag, New York.

[5] Fan, J. (1990). Asymptotic normality for deconvolving kernel density estimators. To appear in *Sankhyā.*

[6] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problem. *Ann. Statist.* To appear.

[7] Fan, J. and Truong, Y. (1990). Nonparametric regression with errors-in-variables. *Institute of Statistics Mimeo Series #2023*, Univ. of North Carolina, Chapel Hill.

[8] Fuller, W. A. (1987). *Measurement error models.* Wiley, New York.

[9] Masry, E. and Rice, J. A. (1991). Gaussian deconvolution via differentiation. To appear in *Can. J. Statist.*.

[10] Prentice, R. L. (1986). Binary regression using an extended Beta-Binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.*, 81, 321–327.

[11] Rice, J. (1984). Bandwidth choice for nonparametric regression. in generalized linear models. *Ann. Statist.*, 12, 1215–1230.

[12] Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika* 72, 583–592.

[13] Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *Ann. Statist.*, 13, 1335–1351.

[14] Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74, 703–716.

[15] Stefanski, L. A. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics*, 21, 169-184.

[16] Whittemore, A.S. and Keller, J.B. (1988). Approximations for errors in variables regression. *J. Amer. Statist. Assoc.*, 83, 1057-1066.

[17] Whittemore, A. S. (1989). Errors-in-variables regression using Stein estimates. *American Statistician.* 43, 226–228.

[18] Zhang, C. H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.*, 18, 806–830.

2038 CHAKRAVARTI, I.M.: A three–class association scheme on the flags of a finite projective plane and a (PBIB) design defined by the incidence of the flags and the Baer subplanes in $PG(2,q^2)$, Oct. 1990. (*8 pages*)

2039 CHAKRAVARTI, I.M.: Geometric construction of some families of two–class and three–class association schemes and codes from non–degenerate and degenerate Hermitian varieties, Oct. 1990. (*13 pages*)

2040 CARLSTEIN, E.: Resampling techniques for stationary time series: Some recent developments, Oct. 1990. (*16 pages*)

2041 FAN, J. & MARRON, S.: Best possible constant for bandwidth selection, Oct. 1990. (*15 pages*)

2042 FAN, J., TRUONG, Y.K. & WANG, Y.: Measurement errors regression: A nonparametric approach, Oct. 1990. (*30 pages*)

2043 REN, J.J.: On Hadamard differentiability and M–estimation in linear models, Sept. 1990. (*dissertation*)