A revisit to correlation analysis for distortion measurement error data

Jun Zhang, Zhenghui Feng, Bu Zhou

Department of Statistics and Applied Probability, National University of Singapore, Singapore

Abstract

In this paper, we consider the estimation problem of a correlation coefficient between unobserved variables of interest. These unobservable variables are distorted in a multiplicative fashion by an observed confounding variable. Two estimators, the moment-based estimator and the direct plug-in estimator, are proposed, and we show their asymptotic normality. Moreover, the direct plug-in estimator is shown asymptotically efficient. Furthermore, we suggest a bootstrap procedure and an empirical likelihood-based statistic to construct the confidence interval. The empirical likelihood statistic is shown to be asymptotically chi-squared. Simulation studies are conducted to examine the performance of the proposed estimators. These methods are applied to analyze the Boston housing price data as an illustration.

Keywords:
Correlation coefficient
Distorting function
Measurement error models
Kernel smoothing

1. Introduction

Measurement error is common in many disciplines, such as economics, health science and medical research, due to improper instrument calibration or many other reasons. Generally, an estimation procedure which ignores measurement error may cause large bias, sometimes seriously large bias. The classical statistical estimation and inference become very challenging. Therefore, it requires particular care to eliminate such bias when estimating target parameters. Research on classical errors-in-variables have been widely studied in the last two decades, for example by Li and Hsiao [12] and Schennach [26] using replicate data, and Carroll et al. [37], Schennach [27] and Wang and Hsiao [3], using instrumental variable methods. Others considered nonparametric or semi-parametric approaches (Carroll et al. [25]; Delaigle et al. [36]; Liang et al. [46]; Liang and Li [17]; Liang and Ren [19]; Liang and Wang [20]; Schafer [16]; Taupin [1]; Zhou and Liang [6]). In addition, Fuller [9] and Carroll et al. [4] give comprehensive reviews containing many parametric and semi-parametric measurement error models.

In this paper, we consider multiplicative effect type errors, namely, distorting measurement errors. Both the response and predictors are unobservable and distorted by general multiplicative effects of some observable confounding variable as

\[
\begin{align*}
\tilde{Y} &= \phi(U)Y, \\
\tilde{X} &= \psi(U)X,
\end{align*}
\]

(1.1)

© 2013 Elsevier Inc. All rights reserved.
where \((Y, X)^\top\) are the unobservable continuous variables of interest, (the superscript \(\tau\) denotes the transpose operator throughout this paper), while \((\tilde{Y}, \tilde{X})^\top\) are available distorted variables, \(\phi(\cdot)\) and \(\psi(\cdot)\) are unknown contaminating functions of an observed confounding variable \(U\), and \(\tilde{U}\) is independent of \((X, Y)^\top\).

The distortion measurement errors model (1.1) usually occurs in biomedical and health-related studies. The confounding variable (for example, it can be the body mass index (BMI), height or weight) usually has some kind of multiplicative effect on the primary variables of interest. Kaysen et al. [11] analyzed the relationship between fibrinogen level and serum transferrin level among hemodialysis patients, and realized that BMI plays the role of confounding variable that may contaminate the fibrinogen level and the serum transferrin level simultaneously. To eliminate the potential bias, Kaysen et al. [11] simply divided the observed fibrinogen level—\(Y\) and observed serum transferrin level—\(\tilde{X}\) by BMI—\(U\). Şentürk and Müller [30] noticed that the exact relationship between the confounding variable (BMI) and primary variables is hardly known in practice. Such a simple way of dividing confounding variable BMI may not be appropriate and may lead to an inconsistent estimator of the target parameters. So Şentürk and Müller [30] proposed model (1.1) as a flexible multiplicative adjustment by involving unknown smooth distorting functions \(\phi(\cdot), \psi(\cdot)\) for the confounding variable \(U\).

To estimate the correlation coefficient between \(Y\) and \(X\), denoted as \(\rho(Y, X)\), Şentürk and Müller [28] observed that the direct calculation of correlation coefficient between \(\tilde{Y}\) and \(\tilde{X}\) will result in an arbitrarily large biased estimator of \(\rho(Y, X)\). Therefore, a proper adjustment method for estimating \(\rho(Y, X)\) needs to be addressed. To solve this problem, Şentürk and Müller [28] established the relationship between \(Y\) and \(X\) through a varying coefficient model, and then employed the binning technique to estimate \(\rho(Y, X)\). Such a transformation procedure can be generalized to regression models with linear structure—for example, linear models [28,30,22,21], generalized linear models [31] and partial linear single index models [43].

The goal of this paper is to construct consistent estimation and do inference for correlation coefficient \(\rho(Y, X)\). Two estimators for \(\rho(Y, X)\) are proposed. One is the moment-based estimator, and the other is the direct plug-in estimator. The basic ideas and motivations of these two methods are summarized as follows.

- **Our first estimator is based on** \(\rho(Y, X) = \rho_0(\epsilon_{YU}, \epsilon_{XU})/\Delta\) for some unknown constant \(\Delta\), where \(\rho_0(\epsilon_{YU}, \epsilon_{XU})\) is the correlation coefficient between \(\epsilon_{YU} = \tilde{Y} - E[Y|U]\) and \(\epsilon_{XU} = \tilde{X} - E[\tilde{X}|U]\), i.e., \(\rho_0(\epsilon_{YU}, \epsilon_{XU}) = \frac{Cov(\epsilon_{YU}, \epsilon_{XU})}{\sqrt{Var(\epsilon_{YU})Var(\epsilon_{XU})}}\) in the population level. This relationship \(\rho(Y, X) = \rho_0(\epsilon_{YU}, \epsilon_{XU})/\Delta\) was first revealed in Appendix D of Şentürk and Müller [28], but they did not study the statistical properties and simulations further. However, \(\rho(Y, X) = \rho_0(\epsilon_{YU}, \epsilon_{XU})/\Delta\) implies that \(\hat{\rho}_0(\epsilon_{YU}, \epsilon_{XU})/\hat{\Delta}\) is also an estimator of \(\rho(Y, X)\), where \(\hat{\rho}_0(\epsilon_{YU}, \epsilon_{XU})\) and \(\hat{\Delta}\) are some estimators of \(\rho_0(\epsilon_{YU}, \epsilon_{XU})\) and \(\Delta\), respectively. So it is worth studying the estimation procedure for \(\hat{\rho}_0(\epsilon_{YU}, \epsilon_{XU})/\hat{\Delta}\) and its associated sample properties. We propose moment-based estimators for \(\rho_0(\epsilon_{YU}, \epsilon_{XU})/\Delta\) and further establish their asymptotic normality properties. Moreover, some simulations are also evaluated. To construct a confidence interval for \(\rho(Y, X)\), a wild bootstrap procedure is proposed.

- **Our second estimator is based on** the recent methodology direct plug-in, proposed by Cui et al. [5]. The direct plug-in method is a convenient tool for distorting measurement error data. The basic idea is to obtain estimators of the distortion functions, say \(\hat{\phi}, \hat{\psi}\) and then calibrate \(Y\) and \(X\) by \(\hat{Y}/\hat{\phi}, \hat{X}/\hat{\psi}\), and finally construct estimation by using these calibrated quantities. The direct plug-in method can be easily adopted in parametric and semi-parametric models, see for instance [5,45,44,13,42]. In this paper, an estimator of \(\rho(Y, X)\) based on the direct plug-in method is also investigated. An interesting result is that the direct plug-in estimator for \(\rho(Y, X)\) is efficient, i.e., the asymptotic variance of the direct plug-in estimator is the same as the classical asymptotic variance of the sample correlation coefficient (see for example [8], Section 8) when data has no distortion effect \(\phi(\cdot) \equiv 1, \psi(\cdot) \equiv 1\). In other words, this direct plug-in estimation procedure for \(\rho(Y, X)\) eliminates the effect caused by the multiplying distorting measurement error \(\phi(U)\), and \(\psi(U)\). Moreover, an empirical likelihood-based statistic is proposed to construct a confidence interval.

Further, we use our proposed estimators to re-analyze Boston housing price data. In [28], the authors used the education level ‘Lstat’ as the confounding variable to investigate the correlation between ‘Crime’ and ‘price’. Zhang et al. [45] indicated that another choice of confounding variable is ‘Pratio’, pupil–teacher ratio by town. We will make a comparison for those estimators of \(\rho(Y, X)\) under these two different choices of confounding variables to see which estimator is more informative and reasonable.

The paper is organized as follows. In Section 2, we propose the moment-based estimator and derive related asymptotic results. A wild bootstrap procedure to construct a confidence interval is also investigated. In Section 3, we give the direct plug-in estimator and present some asymptotic results. We develop a calibrated empirical log-likelihood ratio statistic and show that the statistic has an asymptotic chi-squared distribution. In Section 4, simulation studies are conducted to examine the performance of the proposed methods. In Section 5, the analysis of Boston housing price data is presented. All technical proofs of the asymptotic results are given in the Supplementary material.

### 2. Estimation procedure and asymptotic results

To ensure identifiability for model (1.1), Şentürk and Müller [28] introduced that

\[
E[\phi(U)] = 1, \quad E[\psi(U)] = 1.
\]  
(2.2)
This assumption is similar to that of classical measurement error, for instance, the common assumption $E(e) = 0$ for $W = X + e$, where $W$ is error-prone and $X$ is error-free. Together with (1.1) and (2.2), we have

$$
\begin{align*}
E[Y|U] &= \phi(U)E[Y|U] = \phi(U)E[Y], \\
E[X|U] &= \psi(U)E[X|U] = \psi(U)E[X].
\end{align*}
$$

(2.3)

Define $e_{YU} = \bar{Y} - E[Y|U] = \bar{Y} - \phi(U)E[Y]$ and $e_{XU} = \bar{X} - E[X|U] = \bar{X} - \psi(U)E[X]$. Under (1.1), (2.2) and (2.3), Şentürk and Müller [28] proved that

$$
\rho(e_{YU}, e_{XU}) = \frac{\text{Cov}(e_{YU}, e_{XU})}{\sqrt{\sigma_{e_{YU}}^2 \sigma_{e_{XU}}^2}} = \frac{E[\phi(U)\psi(U)]}{\sqrt{E[\phi^2(U)]E[\psi^2(U)]}} = \rho_{(Y,X)} \Delta,
$$

(2.4)

where $\sigma_{e_{YU}}^2 = \text{Var}(e_{YU})$ and $\sigma_{e_{XU}}^2 = \text{Var}(e_{XU})$—for more details see Appendix D of [28], which indicated that $\Delta$ can be any value in $[0, 1]$ under some mild conditions. From (2.4), Pearson’s correlation coefficient $\rho_{(Y,X)}$ can be directly estimated by $\rho_{(e_{YU}, e_{XU})}/\Delta$ in the population level. Thus, a new estimation procedure can be constructed by using an estimator of Pearson’s correlation coefficient $\rho_{(e_{YU}, e_{XU})}$ between $(e_{YU}, e_{XU})$, together with an estimator of the unknown constant $\Delta$. We describe the procedures in the following subsections.

2.1. Estimating procedure and asymptotic result for $\rho_{(e_{YU}, e_{XU})}$

In (2.4), a moment-based estimation procedure for $\rho_{(e_{YU}, e_{XU})}$ can be constructed by estimating $\text{Cov}(e_{YU}, e_{XU}), \sigma_{e_{YU}}^2$ and $\sigma_{e_{XU}}^2$ respectively. We propose our estimators as

$$
\hat{\text{Cov}}(e_{YU}, e_{XU}) = n^{-1} \sum_{i=1}^{n} \hat{e}_{YUi} \hat{e}_{XUi} - \hat{\bar{Y}} \hat{\bar{X}} = \hat{\text{Cov}}(e_{YU}, e_{XU}),
$$

(2.5)

$$
\hat{\sigma}_{e_{YU}}^2 = n^{-1} \sum_{i=1}^{n} \hat{e}_{YUi}^2 - \hat{\bar{Y}}^2, \quad \hat{\sigma}_{e_{XU}}^2 = n^{-1} \sum_{i=1}^{n} \hat{e}_{XUi}^2 - \hat{\bar{X}}^2,
$$

(2.6)

$$
\hat{\sigma}_{YU}^2 = n^{-1} \sum_{i=1}^{n} \hat{e}_{YUi}^2 - \hat{\bar{Y}}^2, \quad \hat{\sigma}_{XU}^2 = n^{-1} \sum_{i=1}^{n} \hat{e}_{XUi}^2 - \hat{\bar{X}}^2,
$$

(2.7)

where $\hat{\bar{Y}} = n^{-1} \sum_{i=1}^{n} \hat{e}_{YU}$ and $\hat{\bar{X}} = n^{-1} \sum_{i=1}^{n} \hat{e}_{XU}$ with $\hat{e}_{YU} = \bar{Y} - \hat{\bar{Y}}$, $\hat{e}_{XU} = \bar{X} - \hat{\bar{X}}$. $\hat{\bar{Y}}$ and $\hat{\bar{X}}$ are the Nadaraya–Watson estimators for $E(Y|U)$ and $E(X|U)$, which are defined as

$$
\hat{\bar{Y}} = \frac{1}{n^{-1} \sum_{j=1}^{n} K_h(U_j - \hat{U}) \hat{Y}_j}{n^{-1} \sum_{j=1}^{n} K_h(U_j - \hat{U})}, \quad \hat{\bar{X}} = \frac{1}{n^{-1} \sum_{j=1}^{n} K_h(U_j - \hat{U}) \hat{X}_j}{n^{-1} \sum_{j=1}^{n} K_h(U_j - \hat{U})}.
$$

(2.8)

Here $K_h(\cdot) = h^{-1}K(\cdot/h)$, and $K(\cdot)$ denotes a kernel density, $h$ is a positive-valued bandwidth. Using (2.5)–(2.7), the estimator of $\rho_{(e_{YU}, e_{XU})}$ can be defined as

$$
\hat{\rho}_{(e_{YU}, e_{XU})} = \frac{\hat{\text{Cov}}(e_{YU}, e_{XU})}{\hat{\sigma}_{e_{YU}} \hat{\sigma}_{e_{XU}}},
$$

(2.9)

We have the following asymptotic result.

Theorem 1. Under Conditions (A1)–(A4), and (A5)(i) given in Appendix A, we have

$$
\sqrt{n} \left[ \hat{\rho}_{(e_{YU}, e_{XU})} - \rho_{(e_{YU}, e_{XU})} \right] \overset{D}{\rightarrow} N(0, \sigma_0^2),
$$

where the explicit form of $\sigma_0^2$ is given in the Supplementary material.

Remark 1. If $\Delta = 1$, i.e., the Cauchy–Schwarz inequality entails that $P(\phi(U) = c \psi(U)) = 1$ for some nonzero constant $c$, it is easy to see that $c = 1$, due to the identifiability condition (2.2). When $\Delta = 1$, we have $\rho_{(e_{YU}, e_{XU})} = \rho_{(Y,X)}$, which means that $\hat{\rho}_{(e_{YU}, e_{XU})}$ is also a root-$n$ consistent estimator of $\rho_{(Y,X)}$ when $\phi(u) = \psi(u)$. Moreover, the asymptotic variance $\sigma_0^2$ can then be reduced to a simple form $\frac{E[\phi^2(U)]}{E[\phi^2(U)]^2} \sigma_0^2$, where $\sigma_0^2$ is the classical asymptotic variance of the sample correlation.
coefficient (see, for example, [8, Section 8]) when the data are observed without distortion. The distorting factor \( \frac{E[\phi^2(U)]}{E[\psi^2(U)]^2} \) is usually greater than one if \( \phi(u) \neq 1 \), which means that the distorting functions \( \phi, \psi \) increase the asymptotic variance.

**Remark 2.** The condition (A5)(i) shows an under-smoothing approach that is unnecessary to achieve the root-\( n \) consistency of \( \rho(e_{yU}, e_{XU}) \). Technically speaking, the bandwidth \( h \) in Assumption (A5)(i) contains the rate \( n^{-1/5} \) of the 'optimal' bandwidth. Thus, most bandwidth selection methods can be employed here, such as the cross-validation method and the plug-in method.

### 2.2. Estimating procedures and asymptotic results for \( \Delta \) and \( \rho(Y,X) \)

We first propose the estimator for \( \Delta \). Note that

\[
E[\phi(U)\psi(U)] = E[\phi(U) - 1]\{\psi(U) - 1\} + 1,
\]

\[
E[\phi^2(U)] = E[(\phi(U) - 1)^2] + 1,
\]

\[
E[\psi^2(U)] = E[(\psi(U) - 1)^2] + 1.
\]

Moreover, (2.3) and the identifiability condition (2.2) entail that \( \phi(U) = E[Y|U] = E[Y|U] = E[X|U] = E[X|U] = E[\tilde{Y}|U] = E[\tilde{Y}|U] = E[\tilde{X}|U] = E[\tilde{X}|U] \). Directly using (2.8), it is easily seen that

\[
\hat{\phi}(U_i) = \frac{\hat{E}_h(\tilde{Y}_i|U_i)}{\tilde{Y}_i}, \quad \hat{\psi}(U_i) = \frac{\hat{E}_h(\tilde{X}_i|U_i)}{\tilde{X}_i},
\]

where \( \tilde{Y} = \frac{1}{n} \sum_{i=1}^{n} \tilde{Y}_i, \tilde{X} = \frac{1}{n} \sum_{i=1}^{n} \tilde{X}_i \). In (2.12), we choose another bandwidth \( h_1 \), a different bandwidth from \( h \) used in (2.8). This is because, in Theorem 2, under-smoothing for \( h \) is needed such that the bias of \( \hat{\Delta} \) is small and the asymptotic normality of \( \hat{\Delta} \) can be obtained. Together with (2.10)–(2.12), we have

\[
\hat{E}[\phi(U)\psi(U)] = n^{-1} \sum_{i=1}^{n} [\hat{\phi}(U_i) - 1][\hat{\psi}(U_i) - 1] + 1.
\]

\[
\hat{E}[\phi^2(U)] = n^{-1} \sum_{i=1}^{n} [\hat{\phi}(U_i) - 1]^2 + 1,
\]

\[
\hat{E}[\psi^2(U)] = n^{-1} \sum_{i=1}^{n} [\hat{\psi}(U_i) - 1]^2 + 1.
\]

Thus, from (2.12)–(2.14), the estimator of \( \Delta \) can be defined as

\[
\hat{\Delta} = \frac{\hat{E}[\phi(U)\psi(U)]}{\sqrt{\hat{E}[\phi^2(U)]\hat{E}[\psi^2(U)]}}.
\]

Together with (2.9) and (2.15), our target estimator for \( \rho(Y,X) \) can be constructed as

\[
\hat{\rho}(Y,X) = \frac{\hat{\rho}(e_{yU}, e_{XU})}{\hat{\Delta}}.
\]

We have the following asymptotic results.

**Theorem 2.** Under Conditions (A1)–(A4), and (A5)(ii) in Appendix A,

\[
\sqrt{n} \left[ \hat{\Delta} - \Delta \right] \xrightarrow{d} N(0, \sigma_\Delta^2),
\]

where \( \sigma_\Delta^2 = \sigma_1^2 + \sigma_2^2 - \sigma_3^2 \Delta \), and the explicit forms of \( \sigma_i \), \( i = 1, 2, 3 \), are given in the Supplementary material.

**Remark 3.** As noted in Remark 1, if \( \Delta = 1 \), we can use \( \hat{\rho}(e_{yU}, e_{XU}) \) as an estimator of \( \rho(Y,X) \). In this context, the asymptotic variance of \( \hat{\Delta} \) can be reduced to \( \left[ 1 - \frac{1}{E[\phi(U)]} \right]^2 \), and (2.14) can be directly used as a moment-based estimator of \( E[\phi^2(U)] \).

As a result, a test statistic can be constructed as \( \sqrt{n} \hat{\Delta}/\sqrt{\hat{\Delta}} \) to test whether \( \Delta = 1 \) or not.

In (2.4), we may need to check whether \( \Delta \) is zero or not in some cases. If \( \Delta \approx 0 \) to some extent, the estimator (2.16) cannot be used to estimate \( \rho(Y,X) \), as the denominator \( \hat{\Delta} \) in (2.16) is close to zero. Generally, we can use Theorem 2 to construct a test statistic \( \sqrt{n} \hat{\Delta}/\sqrt{\hat{\Delta}} \) to test \( H_0 : \Delta = 0 \), where \( \hat{\Delta} \) is a plug-in consistent estimator of \( \Delta \). Nevertheless, \( \Delta = 0 \) is equivalent to \( E[\psi(U)\phi(U)] = 0 \). **Corollary 1** in the following presents an asymptotic normality for estimator \( \hat{E}[\phi(U)|\psi(U)] \), which can be used to test \( H_0 : \hat{E}[\phi(U)|\psi(U)] = 0 \) (or equivalently \( \Delta = 0 \)). See more discussions in the Supplementary material.
Corollary 1. Under Conditions (A1)—(A4), and (A5) (ii) in Appendix A,。
\[ \sqrt{n} \left[ \hat{\rho} - \rho \right] \Rightarrow N(0, \sigma^2) \]
where the explicit form of \( \sigma^2 \) is given in the Supplementary material.

Next, we present the asymptotic normality of the moment-based estimator \( \hat{\rho} \) (\ref{eq:2.17}):

**Theorem 3.** Under Conditions (A1)—(A5) in Appendix A, we have that
\[ \sqrt{n} \left[ \hat{\rho} - \rho \right] \Rightarrow N(0, \sigma^2) \]
where \( \sigma^2 = \Delta^2 \sigma_0^2 + \Delta^2 \sigma_0^2 - \Gamma_{\rho \xi U} \), \( \sigma_0^2 \) is the asymptotic variance obtained in Theorem 1, \( \sigma^2 \) is the asymptotic variance obtained in Theorem 2, and \( \Gamma_{\rho \xi U} \) is given in the Supplementary material.

Remark 4. From (2.16), it is easily seen that the asymptotic variance of \( \hat{\rho} \) can be obtained from
\[ \Delta^2 \sigma_0^2 = \Delta^2 \sigma_0^2 + \Delta^2 \sigma_0^2 - \Gamma_{\rho \xi U} \]
Thus, the first term in asymptotic variance \( \sigma^2 \), \( \Delta^2 \sigma_0^2 \), can be viewed as the variance from the first stage estimation of \( \rho \), the second term \( \Delta^2 \sigma_0^2 \) is the variance of the second stage estimation of \( \Delta \), and the third one \( \Gamma_{\rho \xi U} \) is the covariance of two-stage estimators.

### 2.3. Constructing bootstrap confidence intervals

In this section, a simple procedure is designed to construct confidence intervals—although we could use \( \hat{\rho} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \) to construct a confidence interval, where \( z_{\alpha/2} \) stands for the \( (1 - \alpha/2) \)th quantile of the standard Gaussian distribution. As the asymptotic variances obtained in Theorems 1 and 3 are very complex, the finite-sample estimator \( \hat{\sigma} \) for \( \sigma \) will certainly have an effect on the behavior of confidence intervals, which may not be precise in finite samples. Another problem is that the normal approximation confidence intervals may produce unreasonable upper or lower bound. For example, if the true value \( \rho \) is 0.9, the upper bound of the interval, \( \hat{\rho} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \), may be possibly beyond 1. To overcome these problems, we propose a wild bootstrap [40,33,7] technique to mimic the distribution of the statistic \( \hat{\rho} \).

**Step 1:** Generate \( B \) times i.i.d. variables \( \xi_{ib}, i = 1, \ldots, n, b = 1, \ldots, B \) with a Bernoulli distribution which respectively take values at 1, 2, with probability 5 \( \xi \). For each \( b \), calculate the arguments \( \hat{\xi}_{ib}^{(b)} = \xi_{ib} - \xi_{ib}^{(b)} \) and
\[
\hat{\rho}^{(b)} = \frac{\text{Cov}(\xi_{ib}^{(b)}, \xi_{ib}^{(b)})}{\sqrt{\text{Var}(\xi_{ib}^{(b)})} \sqrt{\text{Var}(\xi_{ib}^{(b)})}},
\]
(2.17)
where \( \text{Cov}(\xi_{ib}^{(b)}, \xi_{ib}^{(b)}) = n^{-1} \Sigma_{i=1}^{n} \xi_{ib}^{(b)} \xi_{ib}^{(b)} - n^{-1} \Sigma_{i=1}^{n} \xi_{ib}^{(b)} \xi_{ib}^{(b)} - n^{-1} \Sigma_{i=1}^{n} \xi_{ib}^{(b)} \xi_{ib}^{(b)} - n^{-1} \Sigma_{i=1}^{n} \xi_{ib}^{(b)} \xi_{ib}^{(b)} - n^{-1} \Sigma_{i=1}^{n} \xi_{ib}^{(b)} \xi_{ib}^{(b)}.
\]

**Step 2:** Generate another \( B \) times i.i.d. variables \( \xi_{ib}, i = 1, \ldots, n, b = 1, \ldots, B \) with a Bernoulli distribution which respectively take values at 1, 2, with probability 5 \( \xi \). For each \( b \), compute the arguments
\[
\hat{\xi}_{ib}^{(b)} = \frac{\hat{\rho}^{(b)}(\xi_{ib})}{\sqrt{\text{Var}(\xi_{ib}^{(b)})} \sqrt{\text{Var}(\xi_{ib}^{(b)})}},
\]
(2.18)

**Step 3:** For each \( b \), from (2.17) and (2.18), we calculate the bootstrap argument \( \hat{\rho}^{(b)} = \frac{\hat{\rho}^{(b)}(\xi_{ib}^{(b)}, \xi_{ib}^{(b)})}{\Delta^{(b)}} \) and calculate the \( \kappa / 2 \) and \( 1 - \kappa / 2 \) quantiles of the bootstrap \( \hat{\rho}^{(b)} \) as the \( \kappa / 2 \) confidence interval.

A confidence interval built by this wild bootstrap method performs well and the performance is further confirmed numerically in the simulation studies in Section 4.

### 3. An efficient estimation procedure for \( \rho(X,Y) \)

#### 3.1. Direct plug-in estimation procedure

In (2.4), if \( \Delta \approx 0 \), the estimation procedure proposed in Section 2 cannot be implemented directly. Another alternative estimation method for \( \rho(X,Y) \) is the direct plug-in method proposed by Cui et al. [5]. Firstly, the kernel smoothing method
is used to estimate the unknown distortion functions, namely, \( \hat{\psi}(\cdot) \) and \( \hat{\psi}(\cdot) \). Secondly, unobserved \( Y \) and \( X \) are calibrated by \( \hat{Y} = \hat{Y}/\hat{\phi}(U) \) and \( \hat{X} = \hat{X}/\hat{\psi}(U) \). Finally, these calibrated \((\hat{Y}, \hat{X})^T\) can be employed to estimate target parameters. It is worth mentioning that the direct plug-in method can be easily adopted in linear, nonlinear, generalized linear, and semiparametric models, see for example [44,43,13,45,42].

From (2.3), it is seen that \( \phi(U) = \frac{E(\hat{Y}/U)}{E(\hat{Y})} \), \( \psi(U) = \frac{E(\hat{X}/U)}{E(\hat{X})} \). Following the calibrated procedure proposed by Cui et al. [5], using (2.12), the unobserved \((X, Y)\) can be estimated as

\[
\hat{Y}_i = \frac{\hat{Y}_i}{\hat{\phi}(U_i)}, \quad \hat{X}_i = \frac{\hat{X}_i}{\hat{\psi}(U_i)},
\]

and Pearson's correlation coefficient \( \rho_{(Y,X)} \) can be estimated directly by

\[
\hat{\rho}_{(Y,X)} = \frac{\sum_{i=1}^n [\hat{Y}_i - \bar{\hat{Y}}][\hat{X}_i - \bar{\hat{X}}]}{\sqrt{\sum_{i=1}^n [\hat{Y}_i - \bar{\hat{Y}}]^2 \sum_{i=1}^n [\hat{X}_i - \bar{\hat{X}}]^2}}.
\]

where \( \bar{\hat{Y}} = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i \) and \( \bar{\hat{X}} = \frac{1}{n} \sum_{i=1}^n \hat{X}_i \). We have the following asymptotic results.

**Theorem 4.** Under Conditions (A1)–(A4), (A5)(ii) in Appendix A, we have

\[
\sqrt{n} \left[ \hat{\rho}_{(Y,X)} - \rho_{(Y,X)} \right] \xrightarrow{D} N(0, \sigma^2_{\rho}),
\]

where \( \sigma^2_{\rho} \) is the classical asymptotic variance of the sample correlation coefficient (see, for example [8, Section 8]).

**Remark 5.** It is worth noting that the direct plug-in estimation for correlation \( \rho_{(Y,X)} \) is efficient: the asymptotic variance of \( \hat{\rho}_{(Y,X)} \) is the same as the classical asymptotic variance of the sample correlation coefficient. In other words, in estimating target parameter \( \rho_{(Y,X)} \), the direct plug-in estimation eliminates the effect caused by the multiplying distortion measurement error \( \phi(U) \), \( \psi(U) \). This is different from all the existing results in the distorting measurement error literature. Usually, the distorting measurement errors affect the asymptotic variance of those proposed estimators, see for example, the binning techniques proposed by Šentürk and Müller [31,28,30,29], and the direct plug-in estimation procedures proposed by Cui et al. [5], Li et al. [44,43] and Zhang et al. [13,45,42]. However, for correlation coefficient \( \rho_{(Y,X)} \), our direct plug-in estimation procedure performs ideally, and there is no loss when we use estimators \((\hat{Y}_i, \hat{X}_i)\) to substitute unobserved \((Y_i, X_i)\), i.e., the effect of distorting errors vanishes.

### 3.2. Empirical likelihood methodology

The confidence intervals for \( \rho_{(Y,X)} \) based on the normal approximation can be constructed as \( l_{\rho,nOR} = \{\rho_{(Y,X)} : n(\hat{\rho}_{(Y,X)} - \rho_{(Y,X)})^2 / \sigma^2_{\rho} \leq c_\alpha \} \) after we obtain a consistent estimator \( \hat{\sigma}_{\rho}^2 \), by using \([\hat{Y}_i, \hat{X}_i]\) in analogy to (3.2). However, such a direct plug-in estimation procedure always has poor finite-sample behavior, resulting in poor performances of coverage probabilities.

Another popular method to construct confidence intervals is the empirical likelihood (EL) method proposed by Owen [24] and Qin and Lawless [23]. Empirical likelihood has some attractive advantages, such as that it can avoid estimating asymptotic covariances and improve the accuracy of coverage; it can also be easily implemented, automatically studenitized, and widely applied—see, for example, [39,15,14,41,38,47,18,34,35]. Moreover, for the distortion measurement error setting, the existing literature shows that the EL method can improve coverage probabilities of parameters without estimating complicated asymptotic variances of their estimates—see for example, [44,45]. In the following, we make statistical inference based on the EL principle.

Note that the correlation coefficient \( \rho_{(Y,X)} \) can be estimated through the estimating equation

\[
E \left[ \left\{ \frac{Y - \hat{Y}(X)}{\sigma_Y} - \rho_{(Y,X)} \frac{X - \hat{X}(X)}{\sigma_X} \right\} \right] = 0
\]

in the population level. Motivated by this equation, the respective empirical log-likelihood ratio function can be defined as

\[
\ell_n(\rho) = -2 \max \left\{ \sum_{i=1}^n \log(p_i) : p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \left[ \frac{Y_i - \hat{Y}}{\sigma_Y} - \rho \frac{X_i - \hat{X}}{\sigma_X} \right] = 0 \right\},
\]

where

\[
\hat{\sigma}_Y = \sqrt{n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \hat{\sigma}_X = \sqrt{n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2}, \quad \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.
\]

As we know, the true variables \((Y, X)\) are distorted and unobservable. The empirical log-likelihood ratio function \( l_n(\rho) \) cannot be used directly. So here we consider the EL principle in the distorting measurement error setting. Following the direct plug-in idea proposed by Cui et al. [5], we define an auxiliary random variable \( \hat{z}_{n,i}(\rho) \) as

\[
\hat{z}_{n,i}(\rho) = \left[ \frac{\hat{Y}_i - \bar{Y}}{\hat{\sigma}_Y} - \rho \frac{\hat{X}_i - \bar{X}}{\hat{\sigma}_X} \right] \frac{\hat{X}_i - \bar{X}}{\hat{\sigma}_X},
\]
Simulation studies

In this section, we present numerical results to evaluate the performance of the proposed procedures. In the following simulations, the Epanechnikov kernel \( K(t) = 0.75(1 - t^2)_+ \) is used. Note that the 'optimal' bandwidth or order \( n^{-1/5} \) is satisfied by Condition (A5)(i) of Theorem 1 for estimating \( \rho(e_{XY}, e_{UX}) \), and we can use the popular leave-one-out cross-validation to select bandwidth \( h \). To select bandwidth \( h_1 \) in the process of estimating \( \Delta \) and direct plug-in estimation, an under-smoothing bandwidth for \( h_1 \) is needed due to Condition (C7) in Appendix A. Thus, we use the rule of thumb suggested by Silverman [32], \( h_1 \) was chosen as \( \hat{s}_U n^{-1/3} \), where \( \hat{s}_U \) is the sample deviation of \( U \). As suggested by Carroll et al. [2], an ad-hoc but reasonable choice is \( O(n^{-1/5}) \times n^{-2/15} = O(n^{-1/3}) \).

Example. In this simulation, the confounding covariate \( U \) is generated from Uniform (0, 1) and the true unobserved variables \( (X, Y)^T \) are generated from a bivariate normal distribution with mean vector \((4, 4)^T\), \( \sigma_1^2 = \sigma_2^2 = 1 \); the correlation coefficient \( \rho(X,Y) \) is set to be \(-0.9, -0.5, 0.0, 0.5, 0.9 \) separately. We generate 500 realizations and the sample size is chosen as \( n = 200, 400, \) and 600.

Case 1 The distorting functions \( \phi(U) = \psi(U) = \frac{3u^2 + 1}{4}, (\Delta = 1) \).
Case 2 The distorting functions \( \phi(U) = \frac{3u^2 + 1}{4}, \frac{u}{\psi(U)} = 1.5 - U, (\Delta = 0.8790) \).

Simulation results are reported in Tables 1–5. In Case 1, we know that \( \hat{\rho}(e_{XY}, e_{UX}), \hat{\rho}(Y,X) \) and \( \hat{\rho}^f(Y,X) \) are all estimators of \( \rho(X,Y) \). A comparison is made among moment-based estimators \( \hat{\rho}(e_{XY}, e_{UX}), \hat{\rho}(Y,X) \), direct plug-in estimator \( \hat{\rho}^f(Y,X) \) and the binning estimator \( \hat{\rho}_{bin} \) proposed by Şentürk and Müller [28]. The binning number is chosen as 6 in this example. The means and standard errors for each estimate are reported in Table 1.

From Table 1, it is seen that the performance of all four estimators (moment-based estimators \( \hat{\rho}(e_{XY}, e_{UX}), \hat{\rho}(Y,X) \), direct plug-in estimator \( \hat{\rho}^f(Y,X) \) and binning estimator \( \hat{\rho}_{bin} \)) are close to the true value \( \rho(X,Y) \). In Table 2, we further report the mean of squared errors (MSE) \( \sum_{j=1}^{500} (\hat{\rho}_j - \rho(X,Y))^2 / 500 \) for an estimate \( \hat{\rho}_j, s = 1, \ldots, 500 \) in each sample, and the mean of absolute
value (MAE) $\sum_{i=1}^{500} |\hat{\rho}_i - \rho_{(Y,X)}|/500$ with $\hat{\rho}_i = \hat{\rho}_{(Yi', \epsilon_{Xi'})}$, $\hat{\rho}_{(Y,X)}$, $\hat{\rho}_{(Y,X)}^*$ and $\hat{\rho}_{bin}$. From Table 2, the direct plug-in estimate $\hat{\rho}_{(Y,X)}^*$ is better than $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$, $\hat{\rho}_{(Y,X)}$ and $\hat{\rho}_{bin}$. When $\rho_{(Y,X)} \neq 0$, moment-based estimator $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$, $\hat{\rho}_{(Y,X)}$ are the second best, while the binning method estimator $\hat{\rho}_{bin}$ is the second best when $\rho_{(Y,X)} = 0$.

In Case 2, $\Delta = 0.8790$, $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$ is not a consistent estimator of $\rho_{(Y,X)}$ unless $\rho_{(Y,X)} = 0$, and this is because $\rho_{(Yi', \epsilon_{Xi'})} = 0$ holds when $\rho_{(Y,X)} = 0$. From Table 3, it is easily seen that the mean values of $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$ have large bias when $\rho_{(Y,X)} \neq 0$, while after adjustment by $\hat{\Delta}$, $\hat{\rho}_{(Y,X)}$ performs well and can be close to the true value $\rho_{(Y,X)}$. The value of direct plug-in estimator $\hat{\rho}_{(Y,X)}^*$ and binning estimator $\hat{\rho}_{bin}$ are also close to the true value $\rho_{(Y,X)}$. In Table 4, the direct plug-in estimator $\hat{\rho}_{(Y,X)}^*$ is the best. When $|\rho_{(Yi', \epsilon_{Xi'})}| = 0.9$, $\hat{\rho}_{bin}$ is the second best, and $\hat{\rho}_{(Y,X)}$ is the second best in other cases. Moreover, in Tables 1 and 3, we also report the performance for the estimator of $\Delta$ under two cases. The estimator $\hat{\Delta}$ performs well, especially when $\Delta = 1$.

In Table 5, the performance of the 95% confidence interval for $\rho_{(Yi', \epsilon_{Xi'})}$ constructed by bootstrap procedure (BP) and EL method are reported. The sample size is $n = 600$ here. The coverage probabilities based on the BP approach and EL approach are closer to nominal coverage probability 95%. The lengths of the confidence intervals based on the EL procedure are shorter than those based on the BP method. Moreover, the estimator $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$ performs well when $\Delta = 1$ or $\rho_{(Y,X)} = 0$, while the

### Table 2
Simulation study for Case 1 with sample size $n = 400$. The mean squared errors (MSE) and mean of absolute value (MAE), associated with standard errors (SE), for binning estimator $\hat{\rho}_{bin}$, moment-based estimator $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$, $\hat{\rho}_{(Y,X)}$, $\hat{\rho}_{(Y,X)}^*$ and direct plug-in estimator $\hat{\rho}_{(Y,X)}^*$.

<table>
<thead>
<tr>
<th>$\rho_{(Y,X)}$</th>
<th>$\hat{\rho}_{bin}$</th>
<th>SE</th>
<th>$\hat{\rho}<em>{(Yi', \epsilon</em>{Xi'})}$</th>
<th>SE</th>
<th>$\hat{\rho}_{(Y,X)}$</th>
<th>SE</th>
<th>$\hat{\rho}_{(Y,X)}^*$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = -0.90$</td>
<td>0.1531 $\times 10^{-3}$</td>
<td>0.0981 $\times 10^{-3}$</td>
<td>0.0976 $\times 10^{-3}$</td>
<td>0.0931 $\times 10^{-3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = -0.50$</td>
<td>1.7598 $\times 10^{-3}$</td>
<td>1.7425 $\times 10^{-3}$</td>
<td>1.7415 $\times 10^{-3}$</td>
<td>1.4660 $\times 10^{-3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.00$</td>
<td>2.3215 $\times 10^{-3}$</td>
<td>2.3742 $\times 10^{-3}$</td>
<td>2.3737 $\times 10^{-3}$</td>
<td>2.0128 $\times 10^{-3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.50$</td>
<td>1.8802 $\times 10^{-3}$</td>
<td>1.8442 $\times 10^{-3}$</td>
<td>1.8446 $\times 10^{-3}$</td>
<td>1.5726 $\times 10^{-3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.90$</td>
<td>0.1287 $\times 10^{-3}$</td>
<td>0.1098 $\times 10^{-3}$</td>
<td>0.1098 $\times 10^{-3}$</td>
<td>0.0876 $\times 10^{-3}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3
Simulation study for Case 2. The means and standard errors (SE) for binning estimator $\hat{\rho}_{bin}$, moment-based estimator $\hat{\rho}_{(Yi', \epsilon_{Xi'})}$, $\hat{\rho}_{(Y,X)}$, $\hat{\rho}_{(Y,X)}^*$ and $\hat{\Delta}$ and direct plug-in estimator $\hat{\rho}_{(Y,X)}^*$.

<table>
<thead>
<tr>
<th>$\rho_{(Y,X)}$</th>
<th>$\hat{\rho}_{bin}$</th>
<th>SE</th>
<th>$\hat{\rho}<em>{(Yi', \epsilon</em>{Xi'})}$</th>
<th>SE</th>
<th>$\hat{\rho}_{(Y,X)}$</th>
<th>SE</th>
<th>$\hat{\rho}_{(Y,X)}^*$</th>
<th>SE</th>
<th>$\Delta$</th>
<th>$\Delta$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = -0.90$</td>
<td>0.8998</td>
<td>0.0210</td>
<td>0.7950</td>
<td>0.0204</td>
<td>0.9047</td>
<td>0.0272</td>
<td>0.8979</td>
<td>0.0136</td>
<td>0.8790</td>
<td>0.0171</td>
<td></td>
</tr>
<tr>
<td>$\rho = -0.50$</td>
<td>0.4974</td>
<td>0.0625</td>
<td>0.4408</td>
<td>0.0481</td>
<td>0.5006</td>
<td>0.0350</td>
<td>0.5001</td>
<td>0.0528</td>
<td>0.8790</td>
<td>0.0156</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.00$</td>
<td>0.0034</td>
<td>0.0780</td>
<td>0.0056</td>
<td>0.0637</td>
<td>0.0064</td>
<td>0.0721</td>
<td>0.0061</td>
<td>0.0717</td>
<td>0.8790</td>
<td>0.0140</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.50$</td>
<td>0.4946</td>
<td>0.0622</td>
<td>0.4381</td>
<td>0.0501</td>
<td>0.4972</td>
<td>0.0568</td>
<td>0.4958</td>
<td>0.0527</td>
<td>0.8790</td>
<td>0.0109</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.90$</td>
<td>0.8916</td>
<td>0.0209</td>
<td>0.7918</td>
<td>0.0213</td>
<td>0.8981</td>
<td>0.0242</td>
<td>0.8971</td>
<td>0.0388</td>
<td>0.8790</td>
<td>0.0101</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\rho_{(Y,X)}$</th>
<th>$\hat{\rho}_{bin}$</th>
<th>SE</th>
<th>$\hat{\rho}<em>{(Yi', \epsilon</em>{Xi'})}$</th>
<th>SE</th>
<th>$\hat{\rho}_{(Y,X)}$</th>
<th>SE</th>
<th>$\hat{\rho}_{(Y,X)}^*$</th>
<th>SE</th>
<th>$\Delta$</th>
<th>$\Delta$</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = -0.90$</td>
<td>0.8990</td>
<td>0.0108</td>
<td>0.7922</td>
<td>0.0119</td>
<td>0.8989</td>
<td>0.0157</td>
<td>0.8991</td>
<td>0.0078</td>
<td>0.8790</td>
<td>0.0104</td>
<td></td>
</tr>
<tr>
<td>$\rho = -0.50$</td>
<td>0.5029</td>
<td>0.0318</td>
<td>0.4388</td>
<td>0.0302</td>
<td>0.4982</td>
<td>0.0345</td>
<td>0.4984</td>
<td>0.0328</td>
<td>0.8790</td>
<td>0.0092</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.00$</td>
<td>0.0017</td>
<td>0.0409</td>
<td>0.0007</td>
<td>0.0375</td>
<td>0.0008</td>
<td>0.0425</td>
<td>0.0004</td>
<td>0.0422</td>
<td>0.8790</td>
<td>0.0084</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.50$</td>
<td>0.4945</td>
<td>0.0323</td>
<td>0.4395</td>
<td>0.0273</td>
<td>0.4989</td>
<td>0.0310</td>
<td>0.4994</td>
<td>0.0290</td>
<td>0.8790</td>
<td>0.0065</td>
<td></td>
</tr>
<tr>
<td>$\rho = 0.90$</td>
<td>0.8929</td>
<td>0.0103</td>
<td>0.7922</td>
<td>0.0116</td>
<td>0.8992</td>
<td>0.0134</td>
<td>0.8997</td>
<td>0.0079</td>
<td>0.8790</td>
<td>0.0054</td>
<td></td>
</tr>
</tbody>
</table>

Table 4
Simulation study for Case 2 with sample size \(n = 400\). The mean squared errors (MSE) and mean of absolute value (MAE), associated with standard errors (SE), for binning estimator \(\hat{\rho}_{\text{bin}}\), moment-based estimator \(\hat{\rho}_{(Y,X)}\) and direct plug-in estimator \(\hat{\rho}^\circ_{(Y,X)}\).

\[
\begin{array}{cccccc}
\rho & \hat{\rho}_{\text{bin}} & \hat{\rho}_{(Y,X)} & \hat{\rho}^\circ_{(Y,X)} \\
\hline
-0.90 & 0.2653 \times 10^{-3} & 0.3346 \times 10^{-3} & 0.0850 \times 10^{-3} \\
-0.50 & 2.2254 \times 10^{-3} & 4.452 \times 10^{-3} & 1.232 \times 10^{-3} \\
0.00 & 2.6853 \times 10^{-3} & 2.6052 \times 10^{-3} & 2.5904 \times 10^{-3} \\
0.50 & 1.8218 \times 10^{-3} & 1.7166 \times 10^{-3} & 1.6118 \times 10^{-3} \\
0.90 & 0.2097 \times 10^{-3} & 0.2520 \times 10^{-3} & 0.0865 \times 10^{-3} \\
-0.90 & 0.0103 & 0.0147 & 0.0074 \\
-0.50 & 0.0323 & 0.0311 & 0.0298 \\
0.00 & 0.0413 & 0.0407 & 0.0404 \\
0.50 & 0.0340 & 0.0328 & 0.0316 \\
0.90 & 0.0117 & 0.0129 & 0.0074 \\
\end{array}
\]

Table 5
Coverage probability and average length for the 95% confidence interval based on sample size \(n = 600\). ‘Lower’ stands for the lower bound, ‘Upper’ stands for upper bound, ‘AL’ stands for average length, ‘CP’ stands for the coverage probability.

\[
\begin{array}{ccccccccc}
\rho_{(Y,X)} & \hat{\rho}^\circ_{(Y,X)} & \hat{\rho}_{(Y,X)} & \hat{\rho}_{(Y,X)} \\
\hline
\text{Bootstrap} & \text{Bootstrap} & \text{Empirical likelihood} \\
\hline
\text{Lower} & \text{Upper} & \text{AL} & \text{CP} & \text{Lower} & \text{Upper} & \text{AL} & \text{CP} \\
\text{Case 1, } \Delta = 1 \\
-0.90 & -0.9131 & -0.8791 & 0.0340 & 95.2\% & -0.9135 & -0.8794 & 0.0340 & 95.2\% & -0.9138 & -0.8838 & 0.0299 & 95.4\% \\
-0.50 & -0.5604 & -0.4293 & 0.1311 & 95.4\% & -0.5606 & -0.4294 & 0.1312 & 95.4\% & -0.5589 & -0.4419 & 0.1170 & 94.7\% \\
0.00 & 0.0845 & 0.0898 & 0.1743 & 94.4\% & 0.0845 & 0.0898 & 0.1743 & 94.4\% & -0.0791 & 0.0822 & 0.1613 & 95.1\% \\
0.50 & 0.4336 & 0.5643 & 0.1308 & 95.4\% & 0.4336 & 0.5643 & 0.1308 & 95.4\% & 0.4386 & 0.5559 & 0.1173 & 94.8\% \\
0.90 & 0.8817 & 0.9153 & 0.0336 & 94.8\% & 0.8817 & 0.9153 & 0.0336 & 94.8\% & 0.8847 & 0.9143 & 0.0296 & 96.0\% \\
\text{Case 2, } \Delta = 0.8790 \\
-0.90 & -0.8138 & -0.7685 & 0.0453 & 0.0\% & -0.9279 & -0.8703 & 0.0576 & 93.6\% & -0.9139 & -0.8841 & 0.0298 & 95.8\% \\
-0.50 & -0.4938 & -0.3829 & 0.1109 & 92.0\% & -0.5612 & -0.4345 & 0.1267 & 95.2\% & -0.5606 & -0.4432 & 0.1172 & 95.2\% \\
0.00 & -0.0709 & 0.0701 & 0.1410 & 95.8\% & -0.0805 & 0.0796 & 0.1601 & 95.8\% & -0.0806 & 0.0797 & 0.1605 & 95.2\% \\
0.50 & 0.3823 & 0.4934 & 0.1111 & 40.4\% & 0.4337 & 0.5605 & 0.1269 & 94.8\% & 0.4420 & 0.5387 & 0.1167 & 95.2\% \\
0.90 & 0.7684 & 0.8139 & 0.0455 & 0.0\% & 0.6966 & 0.9269 & 0.0574 & 95.8\% & 0.8852 & 0.9149 & 0.0297 & 95.4\% \\
\end{array}
\]

coverage probability is zero for \(|\rho_{(Y,X)}| = 0.9\) when \(\Delta = 0.8790\), which means that \(\rho(\epsilon_{(Y|X)}, \epsilon_{(X|U)})\) can only be an estimator of \(\rho_{(Y,X)}\) when \(\Delta = 0.5\) or \(\rho_{(Y,X)} = 0\).

The simulation shows that when \(\Delta = 0.5\) or \(\rho_{(Y,X)} = 0\), generally, four estimators \(\hat{\rho}(\epsilon_{(Y|X)}, \epsilon_{(X|U)}), \hat{\rho}_{(Y,X)}, \hat{\rho}^\circ_{(Y,X)}\) and \(\hat{\rho}_{\text{bin}}\) perform well. When \(\Delta \neq 0\) or \(\rho_{(Y,X)} \neq 0\), \(\hat{\rho}(\epsilon_{(Y|X)}, \epsilon_{(X|U)})\) has large bias and cannot consistently estimate \(\rho_{(Y,X)}\). The direct plug-in estimator \(\hat{\rho}^\circ_{(Y,X)}\) is the winner among them, and the moment-based estimator \(\hat{\rho}_{(Y,X)}\) and binning method \(\hat{\rho}_{\text{bin}}\) are the second best respectively under different settings. Although the performance of \(\hat{\rho}(\epsilon_{(Y|X)}, \epsilon_{(X|U)}), \hat{\rho}_{(Y,X)}\) and \(\hat{\rho}_{\text{bin}}\) are not better than \(\hat{\rho}^\circ_{(Y,X)}\), they still can be used as the benchmark estimators in practice.

We also conduct some simulations to investigate the performance of 95% coverage probability based on normal approximations \(\hat{\rho}(\epsilon_{(Y|X)}, \epsilon_{(X|U)}) \pm z_{a/2} \cdot \hat{\sigma}_{C}/\sqrt{n}\), \(\hat{\rho}_{(Y,X)} \pm z_{a/2} \cdot \hat{\sigma}_{C}/\sqrt{n}\), \(\rho(X_{U}) \pm z_{a/2} \cdot \hat{\sigma}_{C}/\sqrt{n}\) and \(\rho(Y_{X}) \pm z_{a/2} \cdot \hat{\sigma}_{C}/\sqrt{n}\), where \(\hat{\rho}(\epsilon_{(Y|X)}, \epsilon_{(X|U)})\) and \(\hat{\rho}(\epsilon_{(Y|X)}, \epsilon_{(X|U)})\) are obtained from (3.1) and (2.12). Here \(z_{a/2}\) stands for the \((1 - a/2)\)th quantile of the standard Gaussian distribution. When \(\Delta = 1\), the coverage probabilities for \(|\rho_{(Y,X)}| = 0.9\) are always 100%. The upper bounds for \(\rho_{(Y,X)} = 0.9\) are all more than 1. Similarly, the lower bounds for \(\rho_{(Y,X)} = -0.9\) are all less than -1, which indicates that the finite-sample estimates \(\hat{\rho}_{(Y,X)}, \hat{\rho}^\circ_{(Y,X)}\) are over-estimated for \(\rho_{(Y,X)} = 0.9\). When \(|\rho_{(Y,X)}| = 0.5\), the coverage probabilities are all around 93%, which is less than the nominal coverage probability 95%. When \(\rho_{(Y,X)} = 0\), the coverage probabilities become less than 40%, and all asymptotic variances \(\hat{\sigma}_{C}, \hat{\sigma}_{C}\) are under-estimated. In the case
Fig. 1. The estimated curve of distorting functions of housing price (HP) and crime rate (CR) against confounding variables Lstat and Pratio, associated 95% pointwise confidence intervals (dotted lines).

Table 6

<table>
<thead>
<tr>
<th>Method</th>
<th>Lstat Estimate</th>
<th>Lower</th>
<th>Upper</th>
<th>Pratio Estimate</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\rho}_{\text{bin}}$</td>
<td>$-0.2201$</td>
<td>$-0.3113$</td>
<td>$-0.1289$</td>
<td>$\ast$</td>
<td>$\ast$</td>
<td>$\ast$</td>
</tr>
<tr>
<td>$\hat{\rho}_{(Y,X)}$</td>
<td>$-0.4089$</td>
<td>$-0.4957$</td>
<td>$-0.2031$</td>
<td>$-0.3874$</td>
<td>$-0.2415$</td>
<td>$-0.1098$</td>
</tr>
<tr>
<td>$\hat{\rho}_{*}(Y,X)$</td>
<td>$-0.0579$</td>
<td>$-0.1670$</td>
<td>$0.1098$</td>
<td>$-0.2415$</td>
<td>$-0.3313$</td>
<td>$-0.1667$</td>
</tr>
</tbody>
</table>

$\Delta = 0.8790$, $\hat{\rho}_{(\epsilon^*_Y, \epsilon^*_X)} \pm z_{\alpha/2} \frac{\hat{\sigma}_0}{\sqrt{n}}$ cannot be used, as it is a biased confidence interval. For the other two, the performances are similar to those in the case $\Delta = 1$. These simulation results show that the finite-sample estimators $\hat{\rho}_0$, $\hat{\rho}_1$, and $\hat{\rho}_\rho$ are unstable, which are not recommended for the construction of confidence intervals for correlation coefficient $\rho_{(Y,X)}$.

5. Real data analysis

In this section, we analyze the Boston housing price data from [10]. There are 14 variables in this dataset, which are size, location, environment, and other factors that may affect house price. We are interested in the correlation between crime rate (CR) and house prices (HP). There are two potential confounding variables: ‘education level’ – Lstat used by Şentürk and Müller [28] and ‘pupil–teacher ratio by town’ – Pratio suggested by Zhang et al. [45]. Therefore, to make the analysis more interpretable, it is interesting to see how these estimators $\hat{\rho}_{\text{bin}}$, $\hat{\rho}_{(Y,X)}$, and $\hat{\rho}_{*}(Y,X)$ work for this data. We first present the patterns of $\hat{\phi}(u)$ and $\hat{\psi}(u)$ in Fig. 1 under these two different confounding variables. Four plots indicate that $\phi(u)$ and $\psi(u)$ are not linear, suggesting the distortion effect of the crime rate (CR) and house price (HP). Next, we investigate the estimator of $\Delta$. For comparison, we use the notation $\hat{\Delta}_L$, $\hat{\Delta}_P$ for the estimated $\Delta$ using the two different confounding variables Lstat and Pratio, respectively. The estimators are obtained as $\hat{\Delta}_L = 0.4355$ and $\hat{\Delta}_P = 0.6241$. The final results are summarized in Table 6.

When using Lstat as the confounding variable, the binning estimator $\hat{\rho}_{\text{bin}}$ and moment-based estimator $\hat{\rho}_{(Y,X)}$ show a significant negative relation between HP and CR, while direct plug-in estimator $\hat{\rho}_{*}(Y,X)$ shows an uncorrelated relation between HP and CR. Let $(\hat{Y}, \hat{X})$ stands for estimated HP and CR. We fit a linear model based on remitted covariate $(\hat{Y}_i, \hat{X}_i)_{i=1}^n$ from (3.1) when confounding variable $U$ is chosen as Lstat. The fitted line is $\hat{Y} = 22.3079 - 0.0379X$, which shows a slightly negative relationship between CR and HP. This straight line is displayed in Fig. 2. For illustration, we also fit a local linear smoothing curve (thin and solid) and the 95% pointwise confidence bands. From Fig. 2, the local linear smoothing estimator shows a nonlinear pattern between CR and HP. Moreover, HP increases in the beginning then decreases, and finally increases with CR from zero to 20. It is known that $\rho_{(Y,X)}$ (Pearson’s correlation coefficient) is a measure of the linear correlation.
Fig. 2. Confounding variable—Lstat. The estimated local smoothing curve of estimated crime rate $\hat{X}$ (CR) against estimated housing price $\hat{Y}$ (HP) and the associated 95% pointwise confidence intervals (dotted lines), and a fitted line $\hat{Y} = 22.3079 - 0.0379 \hat{X}$ (the straight line).

Fig. 3. Confounding variable—Prtatio. The estimated local smoothing curve of estimated crime rate $\hat{X}$ (CR) against estimated housing price $\hat{Y}$ (HP) and the associated 95% pointwise confidence intervals (dotted lines), and a fitted line $\hat{Y} = 22.3079 - 0.0379 \hat{X}$ (the straight line).

(dependence) between the two variables $Y$ and $X$. The nonlinear pattern in Fig. 2 implies that the linear measure $\hat{\rho}^*_{(Y,X)}$ may not be appropriate. The 95% confidence interval of $\hat{\rho}^*_{(Y,X)}$ in Table 6 implies that there is no linear correlation between HP and CR, but there may exist a nonlinear correlation as indicated in Fig. 2. That is the reason why the direct plug-in estimator $\hat{\rho}^*_{(Y,X)}$ draws a different conclusion from the moment-based estimator $\hat{\rho}_{(Y,X)}$.

Note that the moment-based estimator $\hat{\rho}_{(Y,X)}$ is derived by the correlation coefficient estimator $\hat{\rho}(\hat{e}_Y, \hat{e}_X)$. It is also interesting to check whether or not the linear correlation measurement $\rho(\hat{e}_Y, \hat{e}_X)$ is appropriate for the choice of confounding variable Lstat. We fit a local linear smoothing curve (thin and solid) based on $\{\hat{e}_Y, \hat{e}_X\}_{i=1}^n$, associated with the 95% pointwise confidence band in Fig. 4. The fitted linear regression of $\hat{e}_Y = -0.4303 - 0.1220 \hat{e}_X$ is also presented in this figure. The local linear smoothing in Fig. 4 implies that the relationship between unobservable $\hat{e}_Y, \hat{e}_X$ is nonlinear, first decreasing and then increasing around zero, but a downward trend in general. This indicates that $\rho(\hat{e}_Y, \hat{e}_X)$ is a reasonable estimator for investigating correlation coefficient of $(\hat{e}_Y, \hat{e}_X)$ to a certain extent.

When using Prtatio as the confounding variable, both moment-based estimator $\hat{\rho}_{(Y,X)}$ and direct plug-in estimator $\hat{\rho}^*_{(Y,X)}$ draw the same conclusion that HP and CR are significant. In Fig. 3, a linear model $\hat{Y} = 23.4062 - 0.3614 \hat{X}$ shows a strong negative relationship between CR and HP. Moreover, a local linear smoothing regression shows that HP and CR also have a
Fig. 4. Confounding variable—Lstat. The estimated local smoothing curve of estimated crime rate residuals $\hat{e}_{YU}$ (CR-Residuals) against estimated housing price residuals $\hat{e}_{YU}$ (HP-Residuals) and the associated 95% pointwise confidence intervals (dotted lines), and a fitted line $\hat{e}_{YU} = -0.4303 - 0.1220\hat{e}_{XU}$ (the straight line).

downward trend. These two observations show that the negative value of $\hat{\rho}_{Y,X}^*$ is proper. Next, we check whether the estimator $\hat{\rho}_{(Y_{iU},X_{iU})}$ is appropriate or not. Again, using $\{\hat{e}_{YU}, \hat{e}_{XU}\}_{i=1}^n$, a local linear smoothing curve of $\hat{e}_{YU}$ against $\hat{e}_{XU}$ associated with its 95% pointwise confidence band is fitted and presented in Fig. 5. A fitted linear regression $\hat{e}_{YU} = -0.0011 - 0.2484\hat{e}_{XU}$ is also presented. Both the linear regression and the local linear smoothing estimator imply that the relationship between unobservable $e_{YU}, e_{XU}$ is monotonic non-increasing, which indicates that the negative value of $\hat{\rho}_{(Y_{iU},X_{iU})}$ is also appropriate, and so is $\hat{\rho}_{Y,X}$.

For binning estimator $\hat{\rho}_{\text{bin}}$, the choice of confounding variable Pratio leads to the unreasonable estimator that $|\hat{\rho}_{\text{bin}}| > 1$ when the binning number is from 2 to 40. This may be because the binning estimator is a weighted-average of the estimated varying coefficient functions and the binning estimation is a less sophisticated nonparametric estimation. Generally, $|\hat{\rho}_{\text{bin}}| \leq 1$ can be guaranteed, however, if the varying coefficient functions are not well estimated by binning estimation, the binning estimator $\hat{\rho}_{\text{bin}}$ fails to draw a meaningful conclusion, as indicated in this analysis.

We aim to investigate the underlying correlation between HP and CR in this example. Under the circumstances, when the confounding variable is chosen as Lstat, $\hat{\rho}_{\text{bin}}$ works [28] but $\hat{\rho}_{Y,X}$ and $\hat{\rho}_{Y,X}^*$ fail as indicated above; If Pratio is used,
The density function of China. The paper is partially supported by the NSFC grant (11101063) and NNSF grant (11201306, 11201499) of China.

Acknowledgments

The authors thank two anonymous referees and the editor for their many constructive suggestions that helped us improve both the presentation and the substance of this article. Zhang’s research is supported by the Natural Science Foundation of Hunan, China. We endeavor to further study. 

Appendix A

We list the conditions needed in the theorems and corollary.

(A1) The density function $f_0(u)$ of the confounding variable $U$ is bounded away from 0 and satisfies the Lipschitz condition of order 1 on $U$, where $U$ is a compact support set of $U$.

(A2) $\phi(u)$, $\psi(u)$ have three bounded and continuous derivatives on $U$. The absolute values of $\phi(u)$ and $\psi(u)$ are greater than a positive constant on $U$. Moreover, $E[\phi^4(U)] < \infty$ and $E[\psi^4(U)] < \infty$.

(A3) $E[Y]$ and $E[X]$, $r = 1, \ldots, p$ are bounded away from 0. Moreover, $E[Y^4] < \infty$ and $E[X^4] < \infty$.

(A4) The kernel function $K(\cdot)$ is a symmetric density function about zero and has bounded derivatives. Furthermore, $K(\cdot)$ satisfies a Lipschitz condition on $U$. $|u|^4K(u)du < \infty$, $j = 1, 2$ and $\int u^jK(u)du \not= 0$.

(A5) As $n \to \infty$, the bandwidths $h$ and $h_1$ satisfy

(i) $h \to 0$, $nh^6 \to 0$, $\frac{nh^2}{\log n} \to \infty$.

(ii) $h_1 \to 0$, $nh_1^4 \to 0$, $\frac{nh_1^2}{\log n} \to \infty$.

Condition (A1) ensures that the density function $f_0(u)$ is positive, which implies that the denominators involved in the nonparametric estimators are bounded away from 0. Condition (A2) is a mild smoothness condition on the involved functions. The absolute values of $\phi(u)$, $\psi(u)$ are above zero on the set $U$, which ensures that the denominators involved in the estimating unknown covariates are not equal to zero. Condition (A3) is necessary in the study of covariate-adjusted models, see [44, 43, 13, 45, 42, 5]. The fourth moment conditions are used to entail that the asymptotic variance of theorems and corollary are all finite. Condition (A4) is commonly imposed in nonparametric regression literature. The Gaussian kernel and quadratic kernel satisfy this condition. Condition (A5) is the bandwidth conditions required for our asymptotic results.

Appendix B. Supplementary material

Supplementary material available at Journal of Multivariate Analysis online includes the technical proofs of Theorems 1–5 and Corollary 1, and the explicit forms of asymptotic variances obtained in Theorems 1–3. Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmva.2013.10.004.

References


