Statistical analysis in Berkson measurement errors

Pei Geng

Department of Mathematics
Illinois State University

August 31, 2018

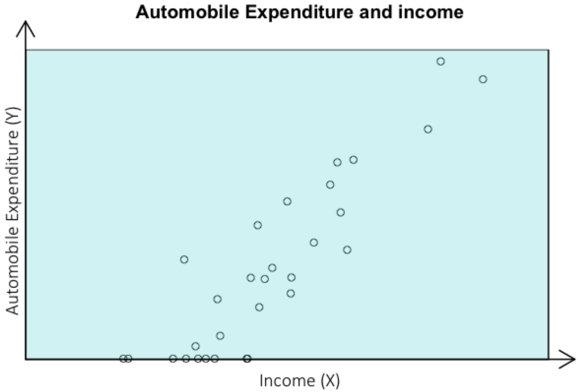
Outline

1. Introduction to measurement error models
   - Errors-in-variables (EIV) models
   - Berkson measurement error models
2. Regression model checking with Berkson measurement errors in covariates using validation data
   - Introduction to the testing problem
   - Parameter estimators and a class of tests based on a minimum distance (m.d.) criterion
   - Main results of the m.d. procedure
   - A finite sample study
3. Ongoing and future work
   - Generalized linear models
   - Varying coefficient autoregressive models

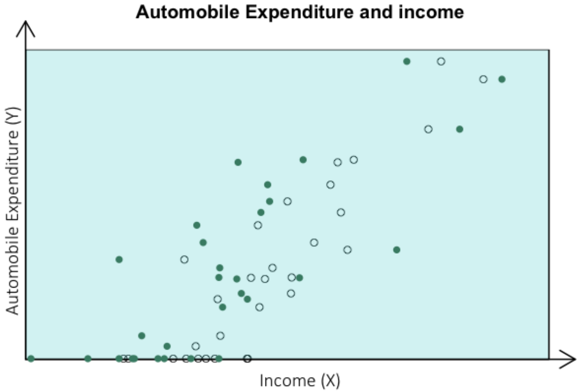
Measurement error models

1. EIV model: $Z = X + u$
   Examples:
   - astronomical data
   - survey or self-reported data: household income, daily calorie intake
2. Berkson model: $X = Z + \eta$
   Examples:
   - oven temperature in chemical experiments
   - lead or air pollutant concentration of a location
3. Literature: Fuller (1987), Cheng and Van Ness (1999), Carroll et al. (2006)
4. Classical methods: Calibration, deconvolution, instrumental variable, validation data.

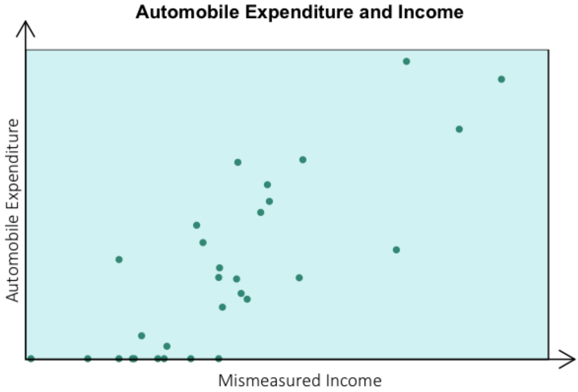
Errors-in-Variables



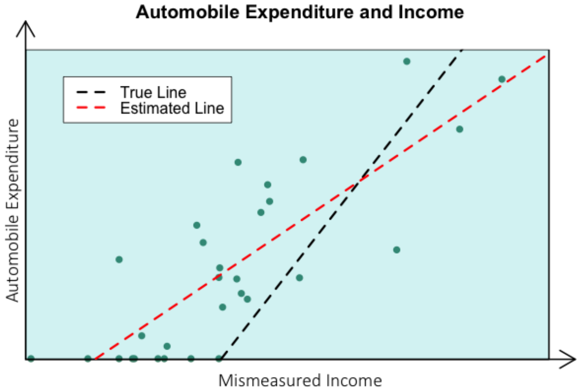
Errors-in-Variables



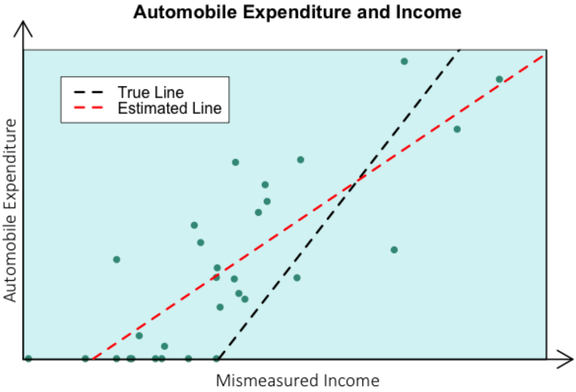
Errors-in-Variables



Errors-in-Variables



Errors-in-Variables



- Measurement errors in covariates mask the pattern of data.
- They cause biased parameter estimation and loss of power in testing.

Berkson models

1. Examples:
   - Income data (Kim, Chao and Härdle (2016))
     The income data were collected by asking individuals which salary range categories they belong to, such as between $5,000 and $9,999, then the midpoint of the range interval $7,500 was used in analysis.
   - Pollutant exposure measurements
     The concentration of atmospheric particulate matter that have a diameter less than 2.5 micrometers (PM2.5) in an area is reported hourly or daily as an average measurement, however, the true exposure for an individual relies on the specific location and the time of the day.
2. Statistical model: $X = Z + \eta$

Testing problem

Regression model:

$$Y = \mu(X) + \varepsilon, \quad X = Z + \eta, \tag{1}$$

where Y is a scalar, X, Z and η are p-dimensional, (ε, Z, η) are mutually independent.

1. Literature:
   - Estimation: Berkson (1950), Huwang and Huang (2000), Wang (2004), Delaigle, Hall and Qiu (2006), Du et al. (2011), Schennach (2013) etc.
   - Hypothesis testing: Koul and Song (2009) (known F_η)
2. We aim to extend the methodology proposed in Koul and Song (2009)(KS) to the case f_η is unknown but when validation data is available.

Testing setup

The problem of interest here is to test

$$H_0 : \mu(x) = m_{\theta_0}(x), \quad \text{for some } \theta_0 \in \Theta \text{ and all } x \in \mathcal{C}, \quad \text{versus}$$
$$H_1 : H_0 \text{ is not true,}$$

based on the primary sample $\{(Z_i, Y_i), i = 1, ..., n\}$ and an independent validation sample $\{(\widetilde{Z}_k, \widetilde{X}_k), k = 1, ..., N\}$, all satisfying (1).
Since X is not observable in primary data, the calibrated regression is obtained as

$$H_\theta(z) := E[m_\theta(X)|Z = z] = \int m_\theta(y + z) f_\eta(y) dy.$$

Example:

1. If $m_\theta(X) = a + bX$, then $H_\theta(Z) = a + bZ$.
2. If $m_\theta(X) = aX^2$, then $H_\theta(Z) = aZ^2 + a\sigma_\eta^2$.
3. In general, the form of H_θ is different from m_θ.

A minimum distance method

Then the original model can be transformed to

$$Y = H(Z) + \xi, \qquad E(\xi|Z) = 0. \tag{2}$$

The hypothesis testing becomes

$$H_0' : H(z) = H_{\theta_0}(z), \text{ for some } \theta_0 \in \Theta \text{ and all } z \in \mathcal{C}, \text{ vs.}$$
$$H_1' : H_0' \text{ is not true.}$$

A minimum distance method

When f_η is known.

- The form of $H_\theta(z)$ is known up to parameter θ.
- The Nadaraya-Watson estimator of regression function is

$$\widehat{H}(z) = \frac{1}{n\hat{f}_w(z)} \sum_{i=1}^{n} K_{hi}(z) Y_i.$$

- Under H_0, the regression function can also be estimated by

$$\widetilde{H}_\theta(z) = \frac{1}{n\hat{f}_w(z)} \sum_{i=1}^{n} K_{hi}(z) H_\theta(Z_i).$$

Nadaraya-Watson estimator

Nadaraya-Watson estimator

A minimum distance method

Koul and Song (AoS, 2009) proposed a m.d. model checking procedure based on the integrated square distance

$$\begin{aligned} M_n(\theta) &= \int_{\mathcal{C}} \left[\widehat{H}(z) - \widetilde{H}_\theta(z)\right]^2 dG(z) \\ &= \int_{\mathcal{C}} \Big[\frac{1}{n\hat{f}_w(z)} \sum_{i=1}^n K_{hi}(z)[Y_i - H_\theta(Z_i)]\Big]^2 dG(z), \\ \tilde{\theta}_n &= \text{argmin}_\theta M_n(\theta). \end{aligned}$$

The asymptotic null distribution:

$$nh^{p/2}\widetilde{\Gamma}_n^{-1/2}\big(M_n(\tilde{\theta}_n) - \widetilde{C}_n\big) \to_d \mathcal{N}_1(0,1).$$

When f_η is unknown. The form of $H_\theta(z)$ is unknown, but the empirical version of η can be obtained by $\widetilde{\eta}_k = \widetilde{X}_k - \widetilde{Z}_k$, $1 \le k \le N$. An estimator of H can be constructed as

$$\widehat{H}_\theta(z) = \frac{1}{N}\sum_{k=1}^{N} m_\theta(z + \widetilde{\eta}_k).$$

The m.d. procedure can be modified as

$$\widehat{M}_n(\theta) = \int_{\mathcal{C}} \Big[\frac{1}{n\hat{f}_w(z)}\sum_{i=1}^{n} K_{hi}(z)[Y_i - \widehat{H}_\theta(Z_i)]\Big]^2 dG(z),$$
$$\hat{\theta}_n = \text{argmin}_\theta \widehat{M}_n(\theta).$$

Then a class of m.d. tests is proposed based on $\widehat{M}_n(\hat{\theta}_n)$.

Assumptions

Define, for $x, y \in \mathbb{R}^p$ and $\theta \in \Theta$,

$$\sigma_\theta(x, y) := \text{Cov}\big(m_\theta(x+\eta), m_\theta(y+\eta)\big), \sigma^2_\theta(x) := \sigma_\theta(x, x).$$

(A1) $\{(Y_i, Z_i), Z_i \in \mathbb{R}^p, i = 1, ..., n\}$ is an i.i.d. sample with regression function $H(z) = E(Y|Z = z)$ satisfying $\int H^2 dG < \infty$, where G is a σ-finite measure with continuous Legesgue density g on $\mathcal{C}$ while $\{(\widetilde{Z}_k, \widetilde{X}_k), \widetilde{Z}_k \in \mathbb{R}^p, \widetilde{X}_k \in \mathbb{R}^p, k = 1, ..., N\}$ is an i.i.d. sample from Berkson measurement error model $X = Z + \eta$.
(A2) $0 < \sigma^2_\varepsilon := Var(\varepsilon) < \infty$, $\tau^2(z) = E[(m_{\theta_0}(X) - H_{\theta_0}(Z))^2 | Z = z]$ is a.e. (G) continuous on $\mathcal{C}$.
(A3) Both $E|\varepsilon|^{2+\delta}$ and $E|(m_{\theta_0}(X) - H_{\theta_0}(Z)|^{2+\delta}$ are finite for some $\delta > 0$.
(A4) Both $E|\varepsilon|^4$ and $E|(m_{\theta_0}(X) - H_{\theta_0}(Z)|^4$ are finite.
(A5) $\int \sigma^2_\theta(z) dG(z) < \infty$, for all $\theta \in \Theta$.

Assumptions contd.

(F1) The density function f_Z is uniformly continuous and bounded away from 0 in $\mathcal{C}$.

(F2) The density function f_Z is twice continuously differentiable in $\mathcal{C}$.

(H1) $m_\theta(x)$ is a.e. continuous in x, for every $\theta \in \Theta$.

(H2) The parametric function family $H_\theta(z)$ is identifiable with respect to θ, i.e, $H_{\theta_1}(z) = H_{\theta_2}(z)$ a.e. in z implies $\theta_1 = \theta_2$.

(H3) For some positive continuous function r on $\mathcal{C}$, and for some $0 < \beta \leq 1$, $|H_{\theta_1}(z) - H_{\theta_2}(z)| \leq \|\theta_1 - \theta_2\|^\beta r(z)$, for all $\theta_1, \theta_2 \in \Theta$ and $z \in \mathcal{C}$.

(H4) For each x, $m_\theta(x)$ is differentiable with respect to θ in a neighborhood of θ_0 with the derivative vector $\dot{m}_\theta(x)$ such that for every sequence $0 < \delta_n \to 0$,

$$\sup_{i,\theta} \frac{\left|\frac{1}{N}\sum_{k=1}^N [m_\theta(Z_i + \widetilde{\eta}_k) - m_{\theta_0}(Z_i + \widetilde{\eta}_k) - (\theta - \theta_0)^T \dot{m}_{\theta_0}(Z_i + \widetilde{\eta}_k)]\right|}{\|\theta - \theta_0\|} = o_p(1),$$

where the supremum is taken over $1 \le i \le n, \|\theta - \theta_0\| \le \delta_n$.
(H5) The vector function $\dot{m}_{\theta_0}(x)$ is continuous in $x \in \mathcal{C}$ and for every $\epsilon > 0$, there are n_ϵ and N_ϵ such that for every $0 < a < \infty$, and for all $n > n_\epsilon, N > N_\epsilon$,

$$P\Big(\max_{1\le i\le n, 1\le k\le N, (nh^p)^{1/2}\|\theta-\theta_0\|\le a} h^{-p/2}\|\dot{m}_\theta(Z_i+\widetilde{\eta}_k)-\dot{m}_{\theta_0}(Z_i+\widetilde{\eta}_k)\| \ge \epsilon\Big) \le \epsilon.$$

(H6) $\int \|\dot{H}_{\theta_0}\|^2 dG < \infty$ and $\Sigma_0 = \int \dot{H}_{\theta_0}\dot{H}_{\theta_0}^T dG$ is positive definite.
(K) The density kernel K is positive symmetric and square integrable on $[-1, 1]^p$.
(W1) $nh^{2p} \to \infty$ and $N/n \to \lambda$, $\lambda > 0$.
(W2) $h \sim n^{-a}$, where $0 < a < \min(1/2p, 4/(p(p+4)))$.

Parameter estimators $\hat{\theta}_n$

Theorem 1

Suppose (A1), (A2), (A5), (F1), (H1)–(H3), (K) and (W1) hold. Then $\hat{\theta}_n \to_p \theta_0$.

Theorem 2

Under H_0, *(A1)–(A3), (A5), (F1)–(F2), (H1)–(H6), (K), (W1)–(W2),*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_q\Big(0, \Sigma_0^{-1}(\Sigma_1 + \lambda^{-1}\Sigma_2)\Sigma_0^{-1}\Big),$$

where Σ_0 *is given in (H6) and*

$$\Sigma_1 = \int \frac{(\sigma_\varepsilon^2 + \tau^2(u))\dot{H}_{\theta_0}(u)\dot{H}_{\theta_0}^T(u)g^2(u)}{f_Z(u)}du,$$

$$\Sigma_2 = \int \sigma_{\theta_0}(x,y)\dot{H}_{\theta_0}(x)\dot{H}_{\theta_0}^T(y)dG(x)dG(y).$$

Interpretation

1. KS showed that $\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d \mathcal{N}_q(0, \Sigma_0 \Sigma_1^{-1} \Sigma_0)$ when f_η is known.
2. Theorem 2 shows that $\hat{\theta}_n$ is $\sqrt{n}$-consistent and the asymptotic covariance matrix is mainly determined by the two terms Σ_1 and Σ_2.
3. The matrix Σ_1 represents the variation in Berkson measurement error model when f_η is known as in KS while Σ_2 represents the contribution due to the estimation of H_θ by $\widehat{H}_\theta$ using the validation data.
4. The covariance tends to decay as N/n increases. When $N/n \to \infty$, in other words, when the validation sample size N is sufficiently large, compared to the primary sample size n, not surprisingly the above asymptotic covariance degenerates to the case as if f_η is known.

Connection between $\hat{\theta}_n$ and $\tilde{\theta}_n$ in linear models

Assume

$$\mu(x) = m_\theta(x) = \theta^T x, \quad x \in \mathcal{C} \subset \mathbb{R}^p, \quad \text{for some } \theta \in \Theta \subset \mathbb{R}^p. \tag{3}$$

(A6) $E\eta^2 < \infty$. $\tau_1(z) := E(|\varepsilon||Z = z)$ is a.e. (G) continuous.
(A7) $\nu_G := \int_{\mathcal{C}} z dG(z) = 0$, $\int_{\mathcal{C}} zz^T dG(z)$ is positive definite.

Proposition 1

Suppose (1) *and* (3) *hold with* $\theta = \theta_0$. *In addition suppose (A1), (F1), (K), (W1), (A6) and (A7) hold, then* $\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) \to_p 0$.

Testing

Theorem 3

Suppose (A1), (A2), (A4), (A5), (F1)–(F2), (K), (H1)–(H6), (W1) and (W2) hold. Then, under H_0,

$$nh^{p/2}\widehat{\Gamma}_n^{-1/2}\left(\widehat{M}_n(\hat{\theta}_n) - \widehat{C}_n\right) \to_d \mathcal{N}_1(0,1),$$

where

$$\hat{\xi}_i = Y_i - \widehat{H}_{\hat{\theta}_n}(Z_i), \quad \widehat{C}_n = \frac{1}{n^2}\sum_{i=1}^{n}\int K_{hi}^2(z)\hat{\xi}_i^2 d\hat{\varphi}(z),$$

$$\widehat{\Gamma}_n = \frac{2h^p}{n^2}\sum_{i \neq j}\left(\int K_{hi}(z)K_{hj}(z)\hat{\xi}_i\hat{\xi}_j d\hat{\varphi}(z)\right)^2.$$

Testing contd.

- Consequently, the null hypothesis is rejected by the test if $\widehat{\mathcal{T}}_n := nh^{p/2}\widehat{\Gamma}_n^{-1/2}|\widehat{M}_n(\hat{\theta}_n) - \widehat{C}_n| > z_{\alpha/2}$ with the asymptotic size $\alpha > 0$.
- Surprisingly, the theorem shows that the sample size ratio N/n does not play a role in the limiting null distribution. This finding is also reflected in the finite sample simulation study through the empirical level and power with different choices of N/n.

Consistency of m.d. tests

Define $\rho(\nu, H_\theta) = \int (\nu - H_\theta)^2 dG, \quad T(\nu) = \text{argmin}_\theta \, \rho(\nu, H_\theta)$.

Theorem 4

Suppose (A1), (A2), (A4), (A5), (F1), (F2), (H3), (K), (W1) and (W2) hold and the alternative hypothesis $H_1 : \mu(x) = m(x), \, x \in \mathcal{C}$ satisfies that $\inf_\theta \rho(H, H_\theta) > 0$ and $T(H)$ is unique. Then $|\mathcal{T}_n| \rightarrow_p \infty$ for any consistent estimator θ_n of $T(H)$.

Power under local alternatives

Let a be a known real-valued function with continuous derivative, $A(z) = E(a(X)|Z = z)$ and $A_2(z) = E([a(X)]^2|Z = z)$, $z \in \mathcal{C}$. Assume

$$\int H_\theta A dG = 0, \qquad \forall\, \theta \in \Theta. \tag{4}$$

We consider a sequence of local alternatives

$$\mathcal{H}_{1,n} : \mu(x) = m_{\theta_0}(x) + b_n\, a(x), \qquad b_n = 1/\sqrt{nh^{p/2}}. \tag{5}$$

Theorem 5

Assume (A1)–(A3), (A5), (F1), (F2), (H1)–(H6), (K), (W1) and (W2) hold. Then under (4) *and* (5), $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d \mathcal{N}_q\left(0, \Sigma_0^{-1}(\Sigma_1 + \lambda^{-1}\Sigma_2)\Sigma_0^{-1}\right)$*, where* Σ_0 *is given in (H6),* Σ_1 *and* Σ_2 *are defined in Theorem 2.*

A finite sample study

- Linear and nonlinear regressions for $p = 1$; linear for $p = 2$.
- $K(u) = 0.75(1 - u^2)I_{(|u| \leq 1)}$ for $p = 1$;
  $K(u) = 0.75^2(1 - u_1^2)(1 - u_2^2)I_{(|u_1| \leq 1, |u_2| \leq 1)}$ for $p = 2$.
- Bandwidth $w = c(\log n/n)^{1/(p+4)}$, $c > 0$. We propose to obtain the optimal w by the unbiased cross-validation criterion, i.e.,

$$\begin{aligned} c_n^* &:= \text{argmin}_{0.1 \leq c \leq 10}\, UCV\Big(c(\log n/n)^{1/(p+4)}\Big), \\ w_{opt} &= c_n^*(\log n/n)^{1/(p+4)}. \end{aligned}$$

  where

$$UCV(w) = \frac{(R(K))^p}{nw^p} + \frac{1}{n(n-1)w^p} \sum_{i \neq j = 1}^{n} (K * K - K)\Big(\frac{Z_i - Z_j}{w}\Big),$$

  with $R(K) = \int K^2(x)dx$ and $K * K(x) = \int K(y)K(x - y)dy$.
- $h = \hat{\sigma}_Z\, n^{-1/3}$ for $p = 1$; $h = n^{-1/4.5}$ for $p = 2$.
- $N/n = 4, 1, 1/4$.

Estimation of θ_0 for $p = 1$

Nonlinear case:

$$m_\theta(x) = e^{\theta x}, \quad \theta_0 = -1, \tag{6}$$

where $\varepsilon \sim \mathcal{N}_1(0, 0.2^2), \eta \sim \mathcal{N}_1(0, 0.2^2), Z \sim U[-1, 1]$.

$N/n = 4$	(n, N)	(60,240)	(100,400)	(200,800)	(300,1200)	(400,1600)
	$\lvert\text{BIAS}(\hat{\theta}_n)\rvert$	0.0010	0.0030	0.0008	0.0017	0.0007
	$\text{RMSE}(\hat{\theta}_n)$	0.0716	0.0552	0.0393	0.0311	0.0274
$N/n = 1$	(n, N)	(60,60)	(100,100)	(200,200)	(300,300)	(400,400)
	$\lvert\text{BIAS}(\hat{\theta}_n)\rvert$	0.0012	0.0036	0.0021	0.0015	0.0009
	$\text{RMSE}(\hat{\theta}_n)$	0.0768	0.0591	0.0424	0.0338	0.0293
$N/n = 1/4$	(n, N)	(60,15)	(100,25)	(200,50)	(300,75)	(400,100)
	$\lvert\text{BIAS}(\hat{\theta}_n)\rvert$	0.0063	0.0048	0.0027	0.00014	0.0008
	$\text{RMSE}(\hat{\theta}_n)$	0.0954	0.0730	0.0503	0.0417	0.0355
$\tilde{\theta}_n$	n	60	100	200	300	400
	$\lvert\text{BIAS}(\tilde{\theta}_n)\rvert$	0.0029	0.0044	0.0012	0.0009	0.0005
	$\text{RMSE}(\tilde{\theta}_n)$	0.0686	0.0552	0.0392	0.0325	0.0264

Table 1 : Performance of $\hat{\theta}_n, \tilde{\theta}_n$ in the nonlinear case (6) with $p = 1$.

Estimation of θ_0 for $p = 2$

$$m_\theta(x) = \theta_1 x_1 + \theta_2 x_2, \quad \theta_0 = (\theta_1, \theta_2)^T = (1, 1)^T. \tag{7}$$

- Z_{i1} and Z_{i2} are generated independently from $U[-1, 1]$
- η_{i1} and η_{i2} are generated from $\mathcal{N}_1(0, 0.1^2)$ and $\mathcal{N}_1(0, 0.2^2)$, respectively.
- ε follows $\mathcal{N}_1(0, 0.2^2)$.

Estimation of θ_0 for $p = 2$

$N/n = 4$	(n, N)	(60,240)	(100,400)	(200,800)	(300,1200)	(400,1600)
	$\lvert\text{BIAS}(\hat{\theta}_{n,1})\rvert$	0.0007	0.0031	0.0007	0.0004	0.0009
	$\text{RMSE}(\hat{\theta}_{n,1})$	0.1069	0.0911	0.0515	0.0434	0.0345
	$\lvert\text{BIAS}(\hat{\theta}_{n,2})\rvert$	0.0020	0.0003	0.0034	0.0020	0.0004
	$\text{RMSE}(\hat{\theta}_{n,2})$	0.1048	0.0863	0.0511	0.0428	0.0356
$N/n = 1$	(n, N)	(60,60)	(100,100)	(200,200)	(300,300)	(400,400)
	$\lvert\text{BIAS}(\hat{\theta}_{n,1})\rvert$	0.0012	0.0032	0.0009	0.0003	0.0009
	$\text{RMSE}(\hat{\theta}_{n,1})$	0.1064	0.0895	0.0516	0.0434	0.0345
	$\lvert\text{BIAS}(\hat{\theta}_{n,2})\rvert$	0.0004	0.0014	0.0032	0.0016	0.0001
	$\text{RMSE}(\hat{\theta}_{n,2})$	0.1049	0.0844	0.0516	0.0427	0.0355
$N/n = 1/4$	(n, N)	(60,15)	(100,25)	(200,50)	(300,75)	(400,100)
	$\lvert\text{BIAS}(\hat{\theta}_{n,1})\rvert$	0.0042	0.0041	0.0015	0.0002	0.0005
	$\text{RMSE}(\hat{\theta}_{n,1})$	0.1073	0.0916	0.0516	0.0435	0.0344
	$\lvert\text{BIAS}(\hat{\theta}_{n,2})\rvert$	0.0040	0.0040	0.0012	0.0002	0.0009
	$\text{RMSE}(\hat{\theta}_{n,2})$	0.1079	0.0882	0.0518	0.0429	0.0357
$\tilde{\theta}_n$	n	60	100	200	300	400
	$\lvert\text{BIAS}(\tilde{\theta}_1)\rvert$	0.0070	0.0005	0.0028	0.0021	0.0011
	$\text{RMSE}(\tilde{\theta}_1)$	0.1162	0.0952	0.0560	0.0497	0.0339
	$\lvert\text{BIAS}(\tilde{\theta}_2)\rvert$	0.0023	0.0006	0.0012	0.0022	0.0002
	$\text{RMSE}(\tilde{\theta}_2)$	0.1086	0.0877	0.0513	0.0438	0.0357

Table 2 : Performance of $\hat{\theta}_n, \tilde{\theta}_n$ in the linear case with $p = 2 = q$

Empirical level and power for $p = 1$

- The nonlinear regression as in (6) is chosen as the null models to obtain the empirical level.
- Three alternative models are chosen to demonstrate the power performance.

$$\text{Model 0: } Y = e^{-X} + \varepsilon.$$
$$\text{Model 1: } Y = e^{-X} - 0.2X^2 + \varepsilon.$$
$$\text{Model 2: } Y = e^{-X} + 0.2\sin(2X) + \varepsilon.$$
$$\text{Model 3: } Y = e^{-X} I_{(X \leq 0.4)} + e^{-0.4} I_{(X > 0.4)} + \varepsilon.$$

Empirical level and power for $p = 1$

$N/n = 4$	(n, N)	(60,240)	(100,400)	(200,800)	(300,1200)	(400,1600)
	Model 0	0.043	0.042	0.048	0.045	0.047
	Model 1	0.153	0.183	0.462	0.724	0.878
	Model 2	0.113	0.196	0.438	0.680	0.866
	Model 3	0.163	0.288	0.689	0.936	0.990
$N/n = 1$	(n, N)	(60,60)	(100,100)	(200,200)	(300,300)	(400,400)
	Model 0	0.043	0.045	0.052	0.044	0.048
	Model 1	0.170	0.199	0.481	0.722	0.861
	Model 2	0.130	0.201	0.437	0.680	0.870
	Model 3	0.187	0.325	0.668	0.922	0.990
$N/n = 1/4$	(n, N)	(60,15)	(100,25)	(200,50)	(300,75)	(400,100)
	Model 0	0.062	0.054	0.059	0.055	0.053
	Model 1	0.185	0.227	0.464	0.724	0.851
	Model 2	0.146	0.217	0.464	0.672	0.856
	Model 3	0.228	0.339	0.690	0.914	0.985
$\widetilde{\mathcal{T}}_n$	n	60	100	200	300	400
	Model 0	0.074	0.060	0.044	0.043	0.055
	Model 1	0.145	0.219	0.469	0.680	0.849
	Model 2	0.144	0.230	0.474	0.705	0.902
	Model 3	0.180	0.291	0.646	0.880	0.986

Table 3 : Empirical levels and powers of $\widehat{\mathcal{T}}_n$ and $\widetilde{\mathcal{T}}_n$ tests for the nonlinear null model

Empirical level and power for $p = 2$

- The linear regression as in (7) is chosen as the null models to obtain the empirical level.
- Three alternative models are chosen to demonstrate the power performance.

With $\theta_0 = (0.5, 1)^T$ and $X = (X_1, X_2)^T$,

$$\begin{aligned}
\text{Model } \emptyset : Y &= \theta_0^T X + \varepsilon, \\
\text{Model I} : Y &= \theta_0^T X + 0.2 X_1 X_2 + \varepsilon, \\
\text{Model II} : Y &= \theta_0^T X + 0.5 \sin(2 X_1 X_2) + \varepsilon, \\
\text{Model III} : Y &= \theta_0^T X I_{(\theta_0^T X \leq 0.5)} + 0.5 I_{(\theta_0^T X > 0.5)} + \varepsilon.
\end{aligned}$$

Empirical level and power for $p = 2$

$N/n = 4$	(n, N)	(60,240)	(100,400)	(200,800)	(300,1200)	(400,1600)
	Model $\emptyset$	0.045	0.038	0.042	0.050	0.048
	Model I	0.205	0.470	0.865	0.968	0.996
	Model II	0.066	0.129	0.303	0.519	0.686
	Model III	0.222	0.488	0.901	0.984	0.997
$N/n = 1$	(n, N)	(60,60)	(100,100)	(200,200)	(300,300)	(400,400)
	Model $\emptyset$	0.048	0.035	0.043	0.053	0.049
	Model I	0.218	0.468	0.859	0.970	0.996
	Model II	0.073	0.128	0.313	0.521	0.688
	Model III	0.223	0.476	0.884	0.979	0.998
$N/n = 1/4$	(n, N)	(60,15)	(100,25)	(200,50)	(300,75)	(400,100)
	Model $\emptyset$	0.060	0.047	0.044	0.056	0.045
	Model I	0.234	0.497	0.883	0.975	0.996
	Model II	0.086	0.159	0.347	0.558	0.716
	Model III	0.242	0.522	0.867	0.971	0.995
$\widetilde{\mathcal{T}}_n$	n	60	100	200	300	400
	Model $\emptyset$	0.042	0.036	0.042	0.056	0.047
	Model I	0.199	0.464	0.869	0.975	0.997
	Model II	0.058	0.124	0.302	0.516	0.690
	Model III	0.212	0.477	0.902	0.984	0.997

Table 4 : Empirical levels and powers of $\widehat{\mathcal{T}}_n$ and $\widetilde{\mathcal{T}}_n$ tests under linear null model,

Summary

- In Berkson measurement error regression, a minimum distance model checking method is adapted when validation data is available.
- The consistency and asymptotic normality of the proposed estimators are derived.
- The limiting distributions of the m.d. tests under the null and certain local alternatives are also established.
- A finite sample study shows reasonable performance of both estimation and test.

Onging and future work

- Logistic regression with EIV models
  Suppose the response Y is a binary variable. The Logistic regression can be used to model the probability of Y and covariate X.

$$
\begin{aligned}
P(Y=1|X) &= \frac{e^{\alpha+\beta^T X}}{1+e^{\alpha+\beta^T X}} \\
Z &= X + u
\end{aligned}
$$

- Time varying coefficient autoregressive models with EIV models

$$
\begin{aligned}
x_t &= f_1(x_{t-d})x_{t-1} + f_2(x_{t-d})x_{t-2} + \dots + f_p(x_{t-d})x_{t-p} + \varepsilon_t \\
z_t &= x_t + u_t
\end{aligned}
$$

Thank you!