A TEST FOR VALIDITY OF LINEAR MULTIPLE ERRORS-IN-VARIABLES MODEL

A. Kukush, H. Navara

Taras Shevchenko National University of Kyiv

Kyiv, 2021

Model specification

Consider a model with errors in variables of the form

$$\begin{aligned} y_i &= \langle c, x_i \rangle + \varepsilon_i, \\ w_i &= x_i + \delta_i, \end{aligned} \tag{1}$$

where

- $\langle c, x_i \rangle$ denotes the Euclidean inner product of c and x_i;
- $\{y_i,\ i = 1, ..., n\}$ are observable response variables;
- $\{x_i \in \mathbb{R}^m,\ i = 1, ..., n\}$ are nonrandom (and unknown) m-dimensional vectors of covariates;
- $c \in \mathbb{R}^m$ is an unknown vector of regression parameters;
- $\{\varepsilon_i,\ i = 1, ..., n\}$ are i.i.d. random variables;
- $\{\delta_i,\ i = 1, ..., n\}$ are i.i.d. random vectors;
- $\{w_i,\ i = 1, ..., n\}$ are observed covariates with errors.

Assumptions on the model

Assumption 1

Random variables ε_i and random vectors δ_i are independent and centered with

$$\mathbb{E}\varepsilon_i^2 = \sigma_\varepsilon^2, \quad \mathrm{Cov}(\delta_i) = \sigma_\delta^2 I_m, \tag{2}$$

where σ_ε^2 and σ_δ^2 are positive constants and I_m stands for the $m \times m$ identity matrix.

Assumption 2

The ratio $\lambda := \frac{\sigma_\delta}{\sigma_\varepsilon}$ is within the interval $[a, A]$, where a and A are known positive numbers.

Assumptions on the model

In order to ensure the convergence in distribution of the regression coefficients estimators and variance estimators, we will need the following technical assumption (hereafter bar means averaging over the index varying from 1 to n, e.g., $\overline{x} = n^{-1}\sum_{i=1}^{n} x_i$):

Assumption 3

(i) For some fixed $\tau > 4$, $\mathbb{E}|\varepsilon|^{\tau} < \infty$ and $\mathbb{E}\|\delta\|^{\tau} < \infty$.

(ii) There exists $\lim_{n\to\infty} \overline{xx^T} =: \mu_2$ and the matrix μ_2 is nonsingular.

(iii) There exists $\lim_{n\to\infty} \overline{x} = \mu_1$ and, for each $1 \le i \le j \le k \le l \le m$, there exist

$$\lim_{n\to\infty} \overline{x(i)x(j)x(k)} =: \mu_3(i,j,k), \qquad \lim_{n\to\infty} \overline{x(i)x(j)x(k)x(l)} =: \mu_4(i,j,k,l),$$

where $x(i)$ denotes the ith coordinate of x.

Assumptions on the model

Assumption 3 (continuied)

(iv) There exists $C > 0$ such that for all $n \geq 1$ it holds $\overline{\|x\|^{\tau}} \leq C$, where $\tau > 4$ is from (i).

(v) At the true point of parameter $\theta = (c^T, \sigma^2) \in \mathbb{R}^{m+1}$, the following matrix is nonsingular:

$$B^{(\lambda)} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[s_\theta(y_i^{(\lambda)}, w_i) s_\theta^T(y_i^{(\lambda)}, w_i) \right].$$

Here $s_\theta = (s_c^T; s_{\sigma^2})^T$ are estimating functions, where

$$s_c : \ \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m \quad \text{and} \quad s_{\sigma^2} : \ \mathbb{R} \times \mathbb{R}^m \times \mathbb{R}^{m+1} \to \mathbb{R},$$

$$s_c(y, w; c) = (y - \langle w, c\rangle)(1 + \|c\|^2)w + (y - \langle c, w\rangle)^2 c,$$

$$s_{\sigma^2}(y, w; \theta) = (y - \langle c, w\rangle)^2 - \sigma^2(1 + \|c\|^2).$$

Goal

Our goal is to test a one-sided compound null hypothesis

$$\mathbb{H}_0 : \quad \sigma_\varepsilon \leq \sigma_0$$

vs. a one-sided compound alternative

$$\mathbb{H}_1 : \quad \sigma_\varepsilon > \sigma_0,$$

where $\sigma_0 > 0$ is a given value.

The case of known λ: reduction to homoscedastic model

Assume that λ from Assumption 2 is known, i.e. $a = A$.
Notice that the heteroscedastic model in consideration can be easily reduced to the homoscedastic one. Indeed, denote $y_i^{(\lambda)} := \lambda y_i$, $c^{(\lambda)} := \lambda c$, $\varepsilon_i^{(\lambda)} = \lambda \varepsilon_i$ and observe that

$$y_i^{(\lambda)} = \left\langle c^{(\lambda)}, x_i \right\rangle + \varepsilon_i^{(\lambda)}, \quad w_i = x_i + \delta_i, \tag{3}$$

$i = 1, ..., n$, with $\mathrm{Var}\left(\varepsilon_i^{(\lambda)}\right) = \lambda^2 \sigma_\varepsilon^2 = \sigma_\delta^2 =: \sigma^2$.

The case of known λ: estimation of regression coefficients

- Denote $z_i := \left(y_i^{(\lambda)}; w_i^T\right)^T$, and define an $(m+1) \times (m+1)$ matrix $Z := \sum_{i=1}^n z_i z_i^T$.
- Let $v = (v_0, ..., v_m)^T$ be the eigenvector of Z that corresponds to the least eigenvalue. It can be shown [Masiuk S. et al (2017)] that $\mathbb{P}(v_0 \neq 0) \to 1$ as $n \to \infty$, so, if n is sufficiently large, one can define

$$\hat{c}^{(\lambda)} = \left(-\frac{v_1}{v_0}, ..., -\frac{v_m}{v_0}\right)^T. \tag{4}$$

- $\hat{c}^{(\lambda)}$ is a consistent estimator of $c^{(\lambda)}$ in the homoscedastic model (3) (see also [Van Huffel S. and Vandewalle J. (1991)] and [Kukush A. and Tsaregorodtsev Ya. (2016)]).
- Thus, we can put $\hat{c} := \frac{\hat{c}^{(\lambda)}}{\lambda}$ and $\hat{c}$ will be a consistent estimator of c.

A consistent estimator of σ_ε^2 can be expressed through the residual sum of squares:

$$\hat{\sigma}_\varepsilon^2 = \frac{\frac{1}{n}\sum_{i=1}^n \left(y_i - \langle \hat{c}, w_i \rangle\right)^2}{1 + \lambda^2 \|\hat{c}\|^2}. \tag{5}$$

Theorem: consistency of estimators

Theorem

Let Assumptions 1–3 hold with $a = A$ in Assumption 2 (which corresponds to the case of λ known). Denote $\theta^{(\lambda)} := (\lambda c^T; \lambda^2 \sigma_\varepsilon^2)^T = \left((c^{(\lambda)})^T; \sigma^2\right)^T$ and let s_c, s_{σ^2} and s_θ be defined as in Assumption 3. Then

1. *The estimating equation $\sum_{i=1}^n s_\theta(y_i^{(\lambda)}, w_i^{(\lambda)}; \theta^{(\lambda)}) = 0$, $\theta^{(\lambda)} \in \mathbb{R}^m \times \mathbb{R}_+$, has a solution with probability tending to 1 as $n \to \infty$ and this solution $\hat{\theta}^{(\lambda)}$ converges to the true value $\theta^{(\lambda)}$ in probability.*

Theorem: asymptotic normality of estimators

Theorem

Let Assumptions 1–3 hold with $a = A$ in Assumption 2 (which corresponds to the case of λ known). Denote $\theta^{(\lambda)} := (\lambda c^T; \lambda^2\sigma_\varepsilon^2)^T = \left((c^{(\lambda)})^T; \sigma^2\right)$ and let s_c, s_{σ^2} and s_θ be defined as in Assumption 3. Then

2. *The solution $\hat{\theta}^{(\lambda)}$ is an asymptotically normal estimator of $\theta^{(\lambda)}$, i.e. $\sqrt{n}\left(\hat{\theta}^{(\lambda)} - \theta^{(\lambda)}\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\theta(\lambda)})$, where*

$$\Sigma_{\theta(\lambda)} = (A^{(\lambda)})^{-1} B^{(\lambda)} (A^{(\lambda)})^{-1}, \tag{6}$$

$$A^{(\lambda)} = -\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial s_\theta(y_i^{(\lambda)}, w_i^{(\lambda)})}{\partial \theta^T}\right],$$

$$B^{(\lambda)} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[s_\theta(y_i^{(\lambda)}, w_i^{(\lambda)}) s_\theta^T(y_i^{(\lambda)}, w_i^{(\lambda)})\right]. \tag{7}$$

Theorem: asymptotic normality of estimators

Theorem

Let Assumptions 1–3 hold with $a = A$ in Assumption 2 (which corresponds to the case of λ known). Denote $\theta^{(\lambda)} := (\lambda c^T; \lambda^2\sigma_\varepsilon^2)^T = \left((c^{(\lambda)})^T; \sigma^2\right)$ and let s_c, s_{σ^2} and s_θ be defined as in Assumption 3. Then

3. *Matrices $A^{(\lambda)}$ and $B^{(\lambda)}$ can be consistently estimated as follows:*

$$\hat{A}^{(\lambda)} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n w_i w_i^T - \hat{\sigma}^2 I_m & \mathcal{O} \\ \mathcal{O} & 1 + \|\hat{c}^{(\lambda)}\|^2 \end{pmatrix},$$

$$\hat{B}^{(\lambda)} = \frac{1}{n}\sum_{i=1}^n s_\theta(y_i^{(\lambda)}, w_i^{(\lambda)}; \hat{\theta}^{(\lambda)}) s_\theta^T(y_i^{(\lambda)}, w_i^{(\lambda)}; \hat{\theta}^{(\lambda)}),$$

where $\hat{\sigma}^2$ and $\hat{c}^{(\lambda)}$ are defined by (5) and (4), and $\hat{\Sigma}_{\theta(\lambda)} = (\hat{A}^{(\lambda)})^{-1}\hat{B}^{(\lambda)}(\hat{A}^{(\lambda)})^{-1} \xrightarrow{\mathbb{P}} \Sigma_{\theta(\lambda)}$ as $n \to \infty$.

Theorem: asymptotic normality of estimators

Theorem

Let Assumptions 1–3 hold with $a = A$ in Assumption 2 (which corresponds to the case of λ known). Denote $\theta^{(\lambda)} := (\lambda c^T; \lambda^2\sigma_\varepsilon^2)^T = \left((c^{(\lambda)})^T; \sigma^2\right)$ and let s_c, s_{σ^2} and s_θ be defined as in Assumption 3. Then

4. *The asymptotic variance of $v_{\sigma^2}^2$ of $\hat{\sigma}^2$ is equal to the lower right entry of $\Sigma_{\theta(\lambda)}$ and can be estimated as*

$$\hat{v}_{\sigma^2}^2 := \frac{1}{(1 + \|\hat{c}^{(\lambda)}\|)^2} \frac{1}{n} \sum_{i=1}^{n} s_{\sigma^2}^2(y_i^{(\lambda)}, w_i; \hat{c}^{(\lambda)}, \hat{\sigma}^2).$$

Hypothesis test

Using results presented above, we can formulate the decision rule for the hypothesis testing problem

$$\begin{aligned} \mathbb{H}_0 : &\quad \sigma_\varepsilon \leq \sigma_0 \\ \mathbb{H}_1 : &\quad \sigma_\varepsilon > \sigma_0, \end{aligned}$$

where $\sigma_0 > 0$ is a given value.

It is easy to see that it is equivalent to the following one:

$$\begin{aligned} \mathbb{H}_0' : &\quad \sigma \leq \lambda\sigma_0, \\ \mathbb{H}_1' : &\quad \sigma > \lambda\sigma_0, \end{aligned}$$

and, given the confidence level $\alpha \in (0, 1)$, we can formulate the decision rule:

reject $\mathbb{H}_0$, if $\hat{\sigma}^2 > \lambda^2\sigma_0^2 + z_\alpha \widehat{se}$, and otherwise, do not reject $\mathbb{H}_0$, where z_α is upper quantile of normal law and $\widehat{se} := \frac{\hat{v}_{\sigma^2}}{\sqrt{n}}$.

Hypothesis test

The asymptotic significance level of the test equals α, or more precisely, Type I error satisfies relations:

$$\lim_{n\to\infty} \mathbb{P}\left(\text{reject } \mathbb{H}_0 \mid \sigma_\varepsilon^2 = \sigma_0^2\right) = \alpha;$$

$$\forall \sigma_1^2 \in (0, \sigma_0^2): \quad \lim_{n\to\infty} \mathbb{P}(\text{reject } \mathbb{H}_0 | \sigma_\varepsilon^2 = \sigma_1^2) = 0.$$

The test is consistent, i.e., for each $\sigma_1^2 > \sigma_0^2$,

$$\lim_{n\to\infty} \mathbb{P}(\text{do not reject } \mathbb{H}_0 | \sigma_\varepsilon^2 = \sigma_1^2) = 0.$$

The case of unknown λ: method

In order to perform testing in the case of $a < A$ in Assumption 2, we will use the following method.

- Introduce $a = \lambda_0 < \lambda_1 < ... < \lambda_K = A$ – uniform partition of $[a, A]$.
- For each λ_j, $j = 0, ..., K$, we perform the procedure described in the previous section of the presentation.
- If at least for one λ_j the hypothesis $\mathbb{H}_0$ is not rejected, then we do not reject $\mathbb{H}_0$.

Simulations

1. Set λ, n, σ_ε, $c = (c_1, c_2)$ (that are assumed to be unknown), the interval $[a, A]$ that contains λ and is assumed to be known, number of points K of the partition at $[a, A]$ and an upper bound σ_0 for the hypothesis.
2. Generate a sample of size n of unobservable regressors $x_1 \sim \mathcal{U}[0, 1]$, $x_2 \sim \mathcal{U}[0, 2.5]$.
3. Generate regression errors $\epsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, $\delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$, $i = 1, ..., n$, where $\sigma_\delta = \lambda\sigma_\varepsilon$.
4. Generate observarble sample $y_i = \langle c, x_i \rangle + \varepsilon_i$, $w_i = x_i + \delta_i$, $i = 1, ..., n$.

Simulations

5. Take uniform partition $a = \lambda_0 < \lambda_1 < ... < \lambda_K = A$ of $[a, A]$ and for each λ_j, $j = 1, ..., K$:
   - 5.1 Make transformation $y_i^{(\lambda_j)} := \lambda_j y_i$, $i = 1, ..., N$, and compute $\hat{c}^{(\lambda_j)}$ as in (4) and put $\hat{c}_j := \frac{\hat{c}^{(\lambda_j)}}{\lambda_j}$.
   - 5.2 For the given λ_j, compute

$$\hat{\sigma}_j^2 := \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(\lambda_j)} - \langle \hat{c}^{(\lambda_j)}, w_i \rangle\right)^2}{1 + \|\hat{c}^{(\lambda_j)}\|^2},$$

$$\widehat{se}_j := \frac{1}{\sqrt{n}} \left(\frac{1}{(1 + \|\hat{c}^{(\lambda_j)}\|)^2} \frac{1}{n} \sum_{i=1}^{n} s_{\sigma^2}^2 (y_i^{(\lambda_j)}, w_i; \hat{c}^{(\lambda_j)}, \hat{\sigma}_j^2) \right)^{\frac{1}{2}}.$$

   - 5.3 Put $\nu_j := \mathbb{1}_{\{\hat{\sigma}_j^2 > \lambda_j^2 \sigma_0^2 + z_\alpha \widehat{se}_j\}}$.
6. Reject $\mathbb{H}_0$ if $\nu_j = 1$ for all $j = 1, ..., K$; otherwise, do not reject $\mathbb{H}_0$.

Simulations

In each simulation, for the given λ, n, σ_ε, $c = (c_1, c_2)$, $[a, A]$, K, σ_0 (see tables below) steps 1–5 were repeated 500 times and the percentage of rejections was computed. The results are presented below.

n	100	500	1000	10000
reject $\mathbb{H}_0, \sigma_0 = \sigma$	0	3.4	6	6.4
reject $\mathbb{H}_0, \sigma_0 = 1.25\sigma$	0	0	0	0
reject $\mathbb{H}_0, \sigma_0 = 0.75\sigma$	96.6	100	100	100

Table: Percentage of $\mathbb{H}_0$ rejections; $\sigma = 0.2$; $\lambda = 3.5$, $a = A = 3.5$

n	100	500	1000	10000
reject $\mathbb{H}_0, \sigma_0 = \sigma$	0	0	0	0
reject $\mathbb{H}_0, \sigma_0 = 1.25\sigma$	0	0	0	0
reject $\mathbb{H}_0, \sigma_0 = 0.75\sigma$	47	99.8	100	100

Table: Percentage of $\mathbb{H}_0$ rejections; $\sigma = 0.2$; $\lambda = 3.5$, $a = 3, A = 4$

Simulations: case of hidden variable

Suppose now that the real model is $y_i = \langle c, x_i \rangle + x_i^h + \varepsilon_i$, $w_i = x_i + \delta_i$, $i = 1, ..., n$. But due to misspecification of the model we use only x_1 and x_2 as explanatory variables. We add noise to the variables x_1 and x_2 as it is described in item 3 above and perform hypothesis check. The results are given below.

n	100	500	1000	10000
reject $\mathbb{H}_0, \sigma_0 = \sigma$	93.6	100	100	100
reject $\mathbb{H}_0, \sigma_0 = 1.25\sigma$	4.6	24.4	50.6	100
reject $\mathbb{H}_0, \sigma_0 = 0.75\sigma$	100	100	100	100

Table: Percentage of $\mathbb{H}_0$ rejections; $\sigma = 0.2$; $\lambda = 3.5$, $a = A$, $x^h \sim \mathcal{U}[0, 12]$

n	100	500	1000	10000
reject $\mathbb{H}_0, \sigma_0 = \sigma$	0.2	0.2	0	0
reject $\mathbb{H}_0, \sigma_0 = 1.25\sigma$	0	0	0	0
reject $\mathbb{H}_0, \sigma_0 = 0.75\sigma$	89	100	100	100

Table: Percentage of $\mathbb{H}_0$ rejections; $\sigma = 0.2$; $\lambda = 3.5$, $a = 3, A = 4$, $x^h \sim \mathcal{U}[0, 12]$

References

[1] *Kukush A., Mandel I.*, Does Regression Approximate the Influence of the Covariates or Just Measurement Errors? A Model Validity Test. ArXiv:1911.07556, 2019.

[2] *Masiuk S. V., Kukush A. G., Shklyar S. V., Chepurny M. I. and Likhtarov I. A. (ed.)*, Radiation Risk Estimation: Based on Measurement Error Models. 2nd ed., de Gruyter, 2017.

[3] *Van Huffel S. and Vandewalle J.*, The Total Least Squares Problem. Frontiers in Applied Mathematics, vol. 9. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1991.

[4] *Kukush A. and Tsaregorodtsev Ya.*, Asymptotic normality of total least squares estimator in a multivariate errors-in-variables model $AX = B$. Mod. Stoch. Theory Appl. 3, N1, 2016, 47-57.

[5] *Carroll R. J., Ruppert D., Stefanski L.A. and Crainiceanu C.*, Measurement Error in Nonlinear Models: A Modern Perspective, 2nd ed. Chapman and Hall / CRC, New York (2006).