Minimax Approach to Errors-in-Variables Linear Models

Yu. Golubev

Workshop on Inverse Problems
Heidelberg 28/10/2016–29/10/2016

This talk deals with the simple (EiV) linear regression model

$$
\begin{aligned}
Y_i &= a + bX_i + \epsilon\xi_i, \\
Z_i &= X_i + \sigma\zeta_i,
\end{aligned}
$$

where

- ξ_i and ζ_i are i.i.d. standard Gaussian random variables;
- $X_i \in \mathbb{R}$ are unknown nuisance variables;
- $\epsilon > 0$ and $\sigma > 0$ are known noise levels.

The goal is to estimate unknown parameters $a, b \in \mathbb{R}$ based on $\{Y_i, Z_i,\ i = 1, \ldots, n\}$.

Adcock, R.J. (1877). Note on the method of least squares. *The Analyst*, 4(6), 183–184.

Adcock, R.J. (1878). A problem in least squares. *The Analyst*, 5(2), 53–54.

$\sigma = 0$

The maximum likelihood (ML) estimate is given by

$$[\hat{a}_n, \hat{b}_n] = \arg\min_{a,b} \Big\{ \sum_{i=1}^{n} (Y_i - a - bZ_i)^2 \Big\}$$

$$\hat{b}_n = \frac{\mathbf{Cov}_n(Y, Z)}{\mathbf{Var}_n(Z)}, \quad \hat{a}_n = \bar{Y}_n - \hat{b}_n \bar{Z}_n,$$

where

$$\mathbf{Cov}_n(Y, Z) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y}_n)(Z_i - \bar{Z}_n),$$

$$\mathbf{Var}_n(Z) = \frac{1}{n} \sum_{i=1}^{n} (Z_i - \bar{Z}_n)^2,$$

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \bar{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

$\sigma = 0$

Statistical properties of $\hat{b}_n$ can easily derived from

$$\mathbf{Cov}_n(Y, Z) \stackrel{\mathbf{P}}{=} b\mathbf{Var}_n(X) + \frac{\epsilon\sqrt{\mathbf{Var}_n(X)}}{\sqrt{n}}\xi_\circ,$$

where $\xi_\circ$ is a standard Gaussian random variable

$$\mathbf{Var}_n(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

So,

$$\sqrt{\frac{n\mathbf{Var}_n(X)}{\epsilon^2}}(\hat{b}_n - b) \stackrel{\mathbf{P}}{=} \xi_\circ.$$

$\sigma = 0$

Theorem

Let $\sigma = 0$ and $\epsilon = \epsilon_n$ is such that

$$\lim_{n\to\infty} \frac{\epsilon_n}{\sqrt{n}} = \epsilon_\circ. \quad \textit{(Large Noise Regime)}$$

Then

$$\lim_{n\to\infty} \inf_{\tilde{b}} \sup_{X:\mathbf{Var}_n(X)>0} \sup_{a,b} \mathbf{Var}_n(X)\mathbf{E}(\tilde{b} - b)^2 = \epsilon_\circ^2,$$

where inf is taken over all estimates of b.

$\sigma > 0$

The main goal in the talk is to extend this simple theorem to

$$\sigma^2 > 0.$$

Casella and Berger wrote that the errors in variables model *"is so fundamentally different from the simple linear regression that it is probably best thought of as a different topic"*.

Casella, G. and Berger, R.L. (1990). *Statistical Inference*, Wadsworth & Brooks, Pacific Grove, CA.

$\sigma > 0$

The ML estimates of a, b are given by

$$\left[\hat{a}_n, \hat{b}_n\right] = \arg\max_{a,b} \max_{X_i} \left\{ -\frac{1}{2\epsilon^2} \sum_{i=1}^{n} (Y_i - a - bX_i)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Z_i - X_i)^2 \right\}.$$

This optimization problem admits a simple analytical solution

$$\begin{aligned} \hat{b}_n =& \frac{W_n}{2} + \operatorname{sign}\left[\mathbf{Cov}_n(Y, Z)\right] \sqrt{\frac{W_n^2}{4} + \rho}, \\ \hat{a}_n =& \bar{Y}_n - \hat{b}_n \bar{Z}_n \end{aligned}$$

where

$$W_n = \frac{\mathbf{Var}_n(Y) - \rho_n \mathbf{Var}_n(Z)}{\mathbf{Cov}_n(Y, Z)}, \quad \rho = \frac{\epsilon^2}{\sigma^2}.$$

So, $\hat{b}_n$ depends on $\mathbf{Var}_n(Y)$, $\mathbf{Cov}_n(Z,Y)$, $\mathbf{Var}_n(Z)$.
With the help of the central limit theorem we get as $n \to \infty$

$$\mathbf{Var}_n(Y) - \epsilon^2 \approx b^2 \mathbf{Var}_n(X) + \frac{\sqrt{\mathbf{Var}_n(X)}}{\sqrt{n}}(2b\epsilon\xi) + \frac{\sqrt{2}\epsilon^2}{\sqrt{n}}\zeta,$$

$$\mathbf{Var}_n(Z) - \sigma^2 \approx \mathbf{Var}_n(X) + \frac{\sqrt{\mathbf{Var}_n(X)}}{\sqrt{n}}(2\sigma\xi') + \frac{\sqrt{2}\sigma^2}{\sqrt{n}}\zeta',$$

$$\mathbf{Cov}_n(Z,Y) \approx b\mathbf{Var}_n(X) + \frac{\sqrt{\mathbf{Var}_n(X)}}{\sqrt{n}}\left(b\epsilon\xi + \sigma\xi'\right) + \frac{\epsilon_n\sigma}{\sqrt{n}}\zeta'',$$

where $\zeta, \zeta', \zeta'', \xi, \xi'$ are independent standard Gaussian random variables.

Theorem

Suppose

$$\lim_{n\to\infty} \mathbf{Var}_n(X) = \theta > 0,$$

and $\sigma = \sigma_n, \ \epsilon = \epsilon_n$ *are such that*

$$\lim_{n\to\infty} \frac{\sigma_n^2}{\sqrt{n}} = \lim_{n\to\infty} \frac{\epsilon_n^2}{\sqrt{n}} = 0.$$

Then

$$\lim_{n\to\infty} \mathbf{P}\left\{ \frac{\hat{b}_n - b}{S_n} \leq x \right\} = \mathbf{P}\Big\{ \xi_\circ \leq x \Big\},$$

where $\xi_\circ$ *is a standard Gaussian random variable and*

$$S_n = \frac{1}{\sqrt{n\theta}} \left[\sigma_n^2 \left(b^2 + \frac{\epsilon_n^2}{\theta} \right) + \epsilon_n^2 \right]^{1/2}.$$

A numerical experiment

For given noise level $\sigma \in [0.3, 0.4]$, we generate 100×5000 replications of the EiV model

$$\begin{aligned} Y_i^k =& a + bX_i^k + \xi_i^k, \\ Z_i^k =& X_i^k + \sigma\zeta_i^k, \end{aligned}$$

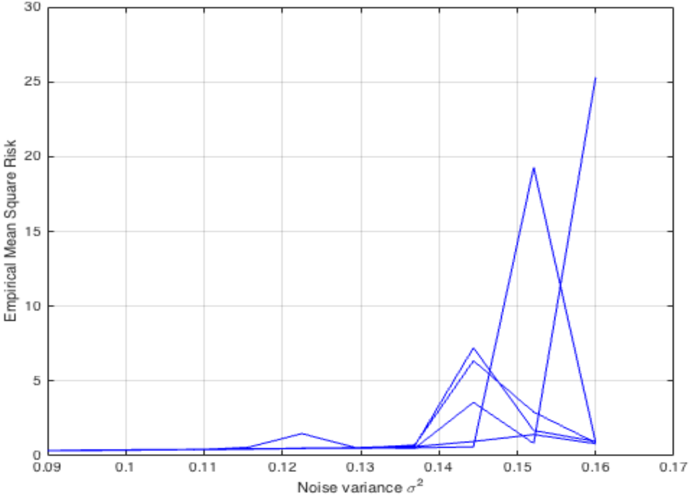
where

- $i = 1, \ldots, 100$, $k = 1, \ldots, 5000$,
- X_i^k are i.i.d. $\mathcal{U}(0, 1)$,
- ξ_i^k and ζ_i^k are i.i.d. $\mathcal{N}(0, 1)$.

We compute the ML estimates $\hat{b}_\sigma^k$ and the empirical mean square risks

$$R(\sigma) = \frac{1}{5000} \sum_{k=1}^{5000} \left[\hat{b}_\sigma^k - b\right]^2.$$

$$\mathrm{E}(\hat{b}_n - b)^2 = \infty$$



Since the ML estimate exhibits an erratic behavior for large σ, we focus on the *Large Noise Regime*, assuming that $\sigma = \sigma_n,\ \epsilon = \epsilon_n$ and

$$\lim_{n\to\infty} \frac{\epsilon_n^2}{\sqrt{n}} = \epsilon_\circ^2 > 0, \quad \lim_{n\to\infty} \frac{\sigma_n^2}{\sqrt{n}} = \sigma_\circ^2 > 0.$$

In this case, we get by the CLT

$$\begin{aligned} \mathbf{Var}_n(Z) - \sigma_n^2 &\approx \mathbf{Var}_n(X) + \sqrt{2}\sigma_\circ^2\zeta, \\ \mathbf{Cov}_n(Y,Z) &\approx b\mathbf{Var}_n(X) + \epsilon_\circ\sigma_\circ\zeta', \\ \mathbf{Var}_n(Y) - \epsilon_n^2 &\approx b^2\mathbf{Var}_n(X) + \sqrt{2}\epsilon_\circ^2\zeta''. \end{aligned}$$

So, our first problem is to estimate b with the help of the following statistics:

$$S_{ZZ} = \mathbf{Var}_n(Z) - \sigma_n^2,$$
$$S_{YY} = \mathbf{Var}_n(Y) - \epsilon_n^2,$$
$$S_{YZ} = \mathbf{Cov}_n(Y, Z).$$

Define nuisance parameter $\theta = \mathbf{Var}_n(X) > 0$.
Then, when n is large, b is estimated based on the following Gaussian observations:

$$S_{ZZ} = \theta + \sqrt{2}\sigma_\circ^2\zeta,$$
$$S_{YY} = b^2\theta + \sqrt{2}\epsilon_\circ^2\zeta',$$
$$S_{YZ} = b\theta + \epsilon_\circ\sigma_\circ\zeta''.$$

This limiting statistical experiment is non-linear.

A simplified model

We begin with estimating b in the following simplified non-linear model:

$$S_{ZZ} = \theta + \sqrt{2}\sigma_\circ^2\zeta,$$
$$S_{YZ} = b\theta + \epsilon_\circ\sigma_\circ\zeta',$$

where ζ, ζ' are i.i.d. standard Gaussian random variables.

In other words, we observe two independent Gaussian random variables

$$X_1 = \mu_1 + \sigma\zeta,$$
$$X_2 = \mu_2 + \sigma'\zeta'$$

and we want to estimate

$$\frac{\mu_1}{\mu_2}.$$

A lower bound

Theorem

$$\liminf_{\delta\to 0} \inf_{\tilde{b}} \sup_{|b-b_\circ|\le\delta/2,\,\theta>0} \frac{\theta^2}{\sigma_\circ^2\epsilon_\circ^2}\mathbf{E}\left[\tilde{b}(S_{YY}, S_{YZ}) - b\right]^2 \ge 1 + 2b_\circ^2\frac{\sigma_\circ^2}{\epsilon_\circ^2},$$

where inf is taken over all estimates of b.

Proof: the Van Trees inequality.

Heuristically, since

$$S_{ZZ} = \theta + \sqrt{2}\sigma_\circ^2\zeta,$$
$$S_{YZ} = b\theta + \epsilon_\circ\sigma_\circ\zeta',$$

when θ is **large**, the optimal estimate of b is

$$\hat{b} = \frac{S_{YZ}}{S_{YY}}$$

and by the Taylor expansion

$$\hat{b} = \frac{b\theta + \epsilon_\circ\sigma_\circ\zeta'}{\theta + \sqrt{2}\sigma_\circ^2\zeta} \approx b + \frac{\sigma_\circ\epsilon_\circ}{\theta}\zeta' - \frac{\sqrt{2}\sigma_\circ^2 b}{\theta}\zeta,$$

we get

$$\mathbf{E}\left[\hat{b} - b\right]^2 \approx \frac{\sigma_\circ^2\epsilon_\circ^2 + 2b\sigma_\circ^4}{\theta^2}.$$

An upper bound

In order to estimate b based on

$$S_{ZZ} = \theta + \sqrt{2}\sigma_\circ^2 \zeta,$$
$$S_{YZ} = b\theta + \epsilon_\circ \sigma_\circ \zeta',$$

we make use of the roughness penalty approach

$$\hat{b}_\alpha = \arg\max_b \max_{\theta>0} \left\{ -\frac{(S_{ZZ}-\theta)^2}{4\sigma_\circ^4} - \frac{(S_{ZY}-b\theta)^2}{2\epsilon_\circ^2\sigma_\circ^2} + \alpha \log(\theta) \right\},$$

where $\alpha > 0$ is a regularization parameter.
This optimization problem admits the following solution:

$$\hat{b}_\alpha = \frac{S_{YZ}}{2\alpha\sigma_\circ^4} \left[\sqrt{\frac{S_{ZZ}^2}{4} + 2\alpha\sigma_\circ^4} - \frac{S_{ZZ}}{2} \right].$$

Theorem

If $\alpha \notin [1.5, 4+\sqrt{7}]$, *then for any* $b \in \mathbb{R}$

$$\max_{\theta>0} \frac{\theta^2}{\sigma_\circ^2 \epsilon_\circ^2} \mathbf{E}(\hat{b}_\alpha - b)^2 > 1 + 2b^2 \frac{\sigma_\circ^2}{\epsilon_\circ^2}.$$

Proof: the high order asymptotic ($\theta \to \infty$) expansion of the risk of $\hat{b}_\alpha$.

Theorem

For any $\alpha \in [1.5, 6.25]$, *uniformly in* $\theta \geq 0$ *and* $b \in \mathbb{R}$

$$\frac{\theta^2}{\sigma_\circ^2 \epsilon_\circ^2} \mathbf{E}(\hat{b}_\alpha - b)^2 \leq 1 + 2b^2 \frac{\sigma_\circ^2}{\epsilon_\circ^2}.$$

Proof: the Monte-Carlo Method.

Complete model

We estimate b based on

$$\begin{aligned} S_{ZZ} &= \theta + \sqrt{2}\sigma_\circ^2\zeta, \\ S_{YY} &= b^2\theta + \sqrt{2}\epsilon_\circ^2\zeta', \\ S_{YZ} &= b\theta + \epsilon_\circ\sigma_\circ\zeta''. \end{aligned}$$

Theorem

$$\liminf_{\delta\to 0} \inf_{\tilde{b}} \sup_{|b-b_\circ|\le\delta,\,\theta>0} \frac{\theta^2}{\sigma_\circ^2\epsilon_\circ^2}\mathbf{E}\left[\tilde{b}(S_{YY}, S_{YZ}, S_{ZZ}) - b\right]^2 \ge 1,$$

where inf is taken over all estimates $\tilde{b}(S_{YY}, S_{YZ}, S_{ZZ})$ of b.

Proof: Van Trees Inequality.

We estimate b as follows:

$$\hat{b}_\alpha = \arg\max_b \left\{ \max_{\theta > 0} L_\alpha(b, \theta; S_{YY}, S_{YZ}, S_{ZZ}) \right\},$$

where

$$L_\alpha(b, \theta; S_{YY}, S_{YZ}, S_{ZZ}) =$$
$$= -\frac{(S_{ZZ} - \theta)^2}{4\sigma_\circ^4} - \frac{(S_{YY} - b^2\theta)^2}{4\epsilon_\circ^4} - \frac{(S_{ZY} - b\theta)^2}{2\epsilon_\circ^2\sigma_\circ^2} + \alpha \log(\theta).$$

There is no formula for $\hat{b}_\alpha$.
However, since $L_\alpha(b, \theta; S_{YY}, S_{YZ}, S_{ZZ})$ is a convex function in $\theta \geq 0$, one can maximize this function with the help of

Coordinate descent algorithm

This method starts with the seed estimate of b

$$\hat{b}_\alpha = \hat{b}_{seed} = \frac{S_{YY}}{\tilde{\theta}_\alpha},$$

$$\tilde{\theta}_\alpha = \frac{S_{YZ}}{2} + \mathrm{sign}(S_{YZ})\sqrt{\frac{S_{YZ}^2}{4} + \alpha\sigma_\circ^2\epsilon_\circ^2}$$

and then

1. update the estimate of θ

$$\hat{\theta}_\alpha = \arg\max_{\theta > 0} L_\alpha(\hat{b}_\alpha, \theta; S_{YY}, S_{YZ}, S_{ZZ})$$

   $\hat{\theta}_\alpha$ is computed analytically;
2. update the estimate of b

$$\hat{b}_\alpha = \arg\max_{b} L_\alpha(b, \hat{\theta}_\alpha; S_{YY}, S_{YZ}, S_{ZZ})$$

   $\hat{b}_\alpha$ is computed analytically;
3. go to 1.

Theorem

$\hat{b}_\alpha$ is minimax for any $\alpha \in [1.5, 2]$, i.e.

$$\sup_{b \in \mathbb{R},\, \theta > 0} \frac{\theta^2}{\sigma_\circ^2 \epsilon_\circ^2} \mathbf{E}\left(\hat{b}_\alpha - b\right)^2 = 1.$$

Proof: the Monte-Carlo Method.

Minimax estimation in the EiV model

We turn to estimating b based on $\{X_i, Y_i, \; i = 1, \ldots, n\}$

$$
\begin{aligned}
Y_i =& a + bX_i + \epsilon_n \xi_i, \\
Z_i =& X_i + \sigma_n \zeta_i.
\end{aligned}
$$

The LNR conditions

$$
\lim_{n\to\infty} \frac{\epsilon_n^2}{\sqrt{n}} = \epsilon_\circ^2 > 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{\sigma_n^2}{\sqrt{n}} = \sigma_\circ^2 > 0.
$$

Theorem

Assume that LNR holds. Then

$$\lim_{n\to\infty} \inf_{\tilde{b}} \sup_{X, b, a} \frac{n\mathbf{Var}_n^2(X)}{\epsilon_n^2 \sigma_n^2} \mathbf{E}\left[\tilde{b}(Y,Z) - b\right]^2 \geq 1,$$

where inf is taken over all estimates $\tilde{b}(Y,Z)$ of b.

Proof: the Van Trees Inequality.

Let

$$\hat{b}_\alpha = \arg\min_b \min_{\theta>0} \left\{ -\frac{[\mathbf{Var}_n(Z) - \sigma_n^2 - \theta]^2}{4\sigma_n^4} - \frac{[\mathbf{Var}_n(Y) - \epsilon_n^2 - b^2\theta]^2}{4\epsilon_n^4} - \frac{[\mathbf{Cov}_n(Y,Z) - b\theta]^2}{2\epsilon_n^2\sigma_n^2} + \frac{\alpha}{n}\log(\theta) \right\}.$$

Theorem

Suppose LNR holds. Then for any $\alpha \in [1.5, 2]$

$$\lim_{n\to\infty} \sup_{X,b,a} \frac{n\mathbf{Var}_n(X)}{\epsilon_n^2\sigma_n^2} \mathbf{E}\left[\hat{b}_\alpha - b\right]^2 = 1.$$

Minimax vs. ML

For given noise level $\sigma \in [0.3, 0.4]$, we generate 100×5000 replications of the EiV model

$$
\begin{aligned}
Y_i^k =& a + bX_i^k + \xi_i^k, \\
Z_i^k =& X_i^k + \sigma\zeta_i^k,
\end{aligned}
$$

where

- $i = 1, \ldots, 100$, $k = 1, \ldots, 5000$,
- X_i^k are i.i.d. $\mathcal{U}(0, 1)$,
- ξ_i^k and ζ_i^k are i.i.d. $\mathcal{N}(0, 1)$.

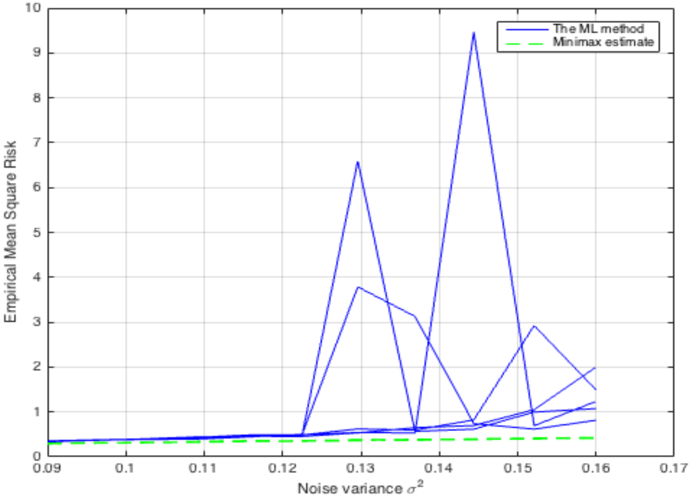
We compute

- the ML estimates $\hat{b}^k$ and the empirical mean square risk

$$R_{ML}(\sigma) = \frac{1}{5000} \sum_{k=1}^{5000} \left[\hat{b}^k - b \right]^2$$

- the minimax estimates $\hat{b}^k_\alpha$ and the empirical mean square risk

$$r_{minmax}(\sigma) = \frac{1}{5000} \sum_{k=1}^{5000} \left[\hat{b}^k_\alpha - b \right]^2.$$



The ML method
Minimax estimate
Empirical Mean Square Risk
Noise variance σ^2
0
1
2
3
4
5
6
7
8
9
10
0.09
0.1
0.11
0.12
0.13
0.14
0.15
0.16
0.17

Summary

- The limiting statistical experiment for the EiV model in Large Noise Regime is non-linear;
- The lower bound for the minimax risk are obtained by the linear approximation of the limit experiment (Van Trees inequality);
- The minimax estimates are computed with the help of the special roughness penalty method.