On the Estimation of Linear Ultrastructural Model when Error Variances are known

Shalabh

Indian Institute of Technology Kanpur (India)

Parameters in a measurement error model can be estimated consistently only when some additional information besides the data set is available.

Various formulations are commonly employed

Interesting formulation - specification of the measurement error variances

Slope parameter in a bivariate model is estimated
- **direct regression** : minimize horizontal distances
- **inverse regression** : minimize vertical distances
- **orthogonal regression** : minimize perpendicular distances from the data points to the regression line

: maximum likelihood estimator under normally distributed measurement errors

Alternative procedures

Received less attention in the literature of measurement error models

- **Technique of reduced major axis:** Estimate slope parameter by geometric mean of direct and inverse regression estimators;
  see, e.g., Sokal and Rohlf (1981)

- Estimate slope parameter by the **arithmetic mean** of direct and inverse regression estimators;
  see, e.g., Aaronson, Bothum, Moutld, Huchra, Schommer and Cornell (1986)

- Slope parameter is estimated by the **slope of the line that bisects the angle** between the direct and inverse regression lines;
  see, e.g., Pierce and Tully (1988).

What is the performance of these estimators under measurement error models?

Which estimator is better under what conditions?

Measurement errors are assumed to be normally distributed

In practice, such an assumption may not always hold true and leads to invalid and erroneous statistical consequences.

What is the effect of departure from normality on the properties of estimators?

Ultrastructural Model

$$\begin{aligned} Y_j &= \alpha + \beta X_j \quad (j = 1, 2, \ldots, n) \\ y_j &= Y_j + u_j \\ x_j &= X_j + v_j \\ X_j &= m_j + w_j \end{aligned}$$

Y_j : true but unobserved values of study variables
X_j : true but unobserved values of explanatory variable
y_j : observed values of study variable
x_j : observed values of explanatory variable
u_j and v_j : associated measurement errors
w_j : random error component
α : unknown intercept term
β : unknown slope parameter

Assume that X_j's are random with possibly different means m_j but same variance

- When $m_1 = m_2 = \ldots = m_n$, ultrastructural model reduces to structural form

- When $w_j = 0 \; \forall \; j = 1, 2, \ldots, n$, ($\sigma_w^2 = 0$), ultrastructural model reduces to functional form

- When $v_j = 0 \; \forall \; j = 1, 2, \ldots, n$, ($\sigma_v^2 = 0$) [i.e., no measurement errors in x_j], ultrastructural model reduces to classical regression model.

Ultrastructural Model: very general framework for measurement error modelling

Distributional assumptions :

- $u_1, u_2, \ldots, u_n \overset{iid}{\sim} (0,\ \sigma_u^2,\ \gamma_{1u}\sigma_u^3,\ (\gamma_{2u}+3)\sigma_u^4)$
- $v_1, v_2, \ldots, v_n \overset{iid}{\sim} (0,\ \sigma_v^2,\ \gamma_{1v}\sigma_v^3,\ (\gamma_{2v}+3)\sigma_v^4)$
- $w_1, w_2, \ldots, w_n \overset{iid}{\sim} (0,\ \sigma_w^2,\ \gamma_{1w}\sigma_w^3\ ,(\gamma_{2w}+3)\sigma_w^4)$
- $u,\ v$ and w are stochastically independent.

γ_1 and γ_2 : Coefficient of skewness and kurtosis respectively

No assumption about the distributional form.

Consistent estimation of α and β with n observations is possible only when some additional information is available

✠ Measurement error variance σ_v^2 is known

$$b_d = \frac{s_{xy}}{s_{xx} - \sigma_v^2} \quad ; s_{xx} > \sigma_v^2$$

where

$$s_{xx} = \frac{1}{n}\sum(x_j - \bar{x})^2 \quad , \quad \bar{x} = \frac{1}{n}\sum x_j \quad ;$$
$$s_{xy} = \frac{1}{n}\sum(x_j - \bar{x})(y_j - \bar{y}) \quad , \quad \bar{y} = \frac{1}{n}\sum y_j$$

Direct regression estimator of slope parameter in the regression of y_j on x_j^* instead of x_j where

$$x_j^* = \bar{x} + \left(1 - \frac{\sigma_v^2}{s_{xx}}\right)(x_j - \bar{x})$$

✠ **Measurement error variance σ_u^2 is known**

$$b_i = \frac{s_{yy} - \sigma_u^2}{s_{xy}} \quad ; s_{yy} > \sigma_u^2$$

where

$$s_{yy} = \frac{1}{n}\sum (y_j - \bar{y})^2$$

Inverse regression estimator b_i essentially arises from the regression of x_j on y_j^* instead of y_j where

$$y_j^* = \bar{y} + \left(1 - \frac{\sigma_u^2}{s_{yy}}\right)(y_j - \bar{y}),$$

b_d and b_i utilizes the knowledge of only one error variance at a time

✠ **Estimator using the knowledge of both the error variances**

$$b_p = t_p + \left(t_p^2 + \frac{\sigma_u^2}{\sigma_v^2}\right)^{\frac{1}{2}} \quad ; \quad s_{xy} \neq 0$$

where

$$t_p = \frac{1}{2s_{xy}}\left(s_{yy} - \frac{\sigma_u^2}{\sigma_v^2}s_{xx}\right)$$

Obtained by orthogonal regression

✠ Estimate β by **reduced major axis**

$$b_r = sign(s_{xy})\,(b_d b_i)^{\frac{1}{2}}$$

: **Geometric mean** of b_d and b_i; [Sokal and Rohlf (1981)]

✠ Estimate β by

$$b_m = \frac{1}{2}\,(b_d + b_i)$$

Mean of estimators b_d **and** b_i; [Aaronson et al. (1986)].

✠ Estimate β by

$$b_b = t_b + (t_b^2 + 1)^{\frac{1}{2}} \quad ; \quad t_b = \frac{b_d b_i - 1}{b_d + b_i}$$

: **Slope of line that bisects the angle between the two regression lines specified by** b_d **and** b_i; [Pierce and Tully (1988)].

Consistency of estimators :

Assume the variance of $m_1, m_2, \ldots, m_n$ tends to a finite quantity σ_m^2 as n tends to infinity.

$$\begin{aligned}
\text{plim}(b_d - \beta) &= 0 \\
\text{plim}(b_i - \beta) &= 0 \\
\text{plim}(b_p - \beta) &= 0 \\
\text{plim}(b_r - \beta) &= 0 \\
\text{plim}(b_m - \beta) &= 0 \\
\text{plim}(b_b - \beta) &= 0
\end{aligned}$$

b_d, b_i, b_p, b_r, b_m and b_b are consistent for β.

Reliability ratios associated with study and explanatory variables are easily available or can be well estimated [Gleser (1992, 1993)].

Reliability ratio = Ratio of variances of true and observed values

Reliability ratios :

$$\lambda_x = \frac{\sigma_m^2 + \sigma_w^2}{\sigma_m^2 + \sigma_w^2 + \sigma_v^2} \; ; \; 0 \leq \lambda_x \leq 1$$
$$\lambda_y = \frac{\beta^2(\sigma_m^2 + \sigma_w^2)}{\beta^2(\sigma_m^2 + \sigma_w^2) + \sigma_u^2} \; ; \; 0 \leq \lambda_y \leq 1$$

Express efficiency properties of estimators as a function of reliability ratios

Helps in obtaining the conditions for the superiority of one estimator over the other in terms of reliability ratios only.

Asymptotic relative variances

$$AsyRelVar(b_d) = \frac{1-\lambda_x}{\lambda_x^2}[\lambda_x + q + (1-\lambda_x)(2+\gamma_{2v})]$$

$$AsyRelVar(b_i) = \frac{1-\lambda_x}{\lambda_x^2}\left[\lambda_x + q + q^2(1-\lambda_x)(2+\gamma_{2u})\right]$$

$$AsyRelVar(b_p) = \frac{1-\lambda_x}{\lambda_x^2}\left[\lambda_x + q + \frac{q^2(1-\lambda_x)}{(q+1)^2}(\gamma_{2u}+\gamma_{2v})\right]$$

$$AsyRelVar(b_r) = \frac{(1-\lambda_x)\delta}{4\lambda_x^2}$$

$$AsyRelVar(b_m) = \frac{(1-\lambda_x)\delta}{4\lambda_x^2}$$

$$AsyRelVar(b_b) = \frac{(1-\lambda_x)\delta}{\lambda_x^2}$$

where

$q = \dfrac{\lambda_x(1-\lambda_y)}{\lambda_y(1-\lambda_x)}$ and $\delta = 2[(1+\lambda_x)(1+q)+(1-\lambda_x)q^2]+(1-\lambda_x)(\gamma_{2v}+q^2\gamma_{2u})$

- Skewness of the distributions of measurement errors has no influence on the asymptotic variances of the estimators.

- Only the kurtosis shows its effect.

- Asymptotic variance for each estimator under normality of errors could be quite different when the distributions depart from normality.

- b_r and b_m are equally efficient.

- b_r and b_m are less efficient than b_p when

$$(q-1)[(3q+1)\gamma_{2v} - q^2(q+3)\gamma_{2u}] < 2(q+1)^2(1-q+q^2).$$

  - When $q = 1$, i.e., $\lambda_x = \lambda_y$, this condition is satisfied for all kinds of error distributions.
  - If $\lambda_x \neq \lambda_y$, the condition invariably holds true provided that both the distributions of measurement errors are mesokurtic or normal.
  - Result remains true for non-normal distributions too when $\gamma_{2v} < 0$ (leptokurtic) and $\gamma_{2u} > 0$ (platykurtic).
  - Both b_r and b_m are asymptotically more efficient than b_p when the condition holds with an opposite inequality sign.

- b_r and b_m are always better than b_b.

- When both the error distributions are assumed to be normal, then b_p is always more efficient than b_r, b_m and b_b, as expected.

Monte Carlo Simulation:

To have an idea about the effect of departure from normal distribution on the efficiency properties of the estimators-

1. normal distribution

2. skew-Normal distribution

3. t distribution

4. beta distribution and

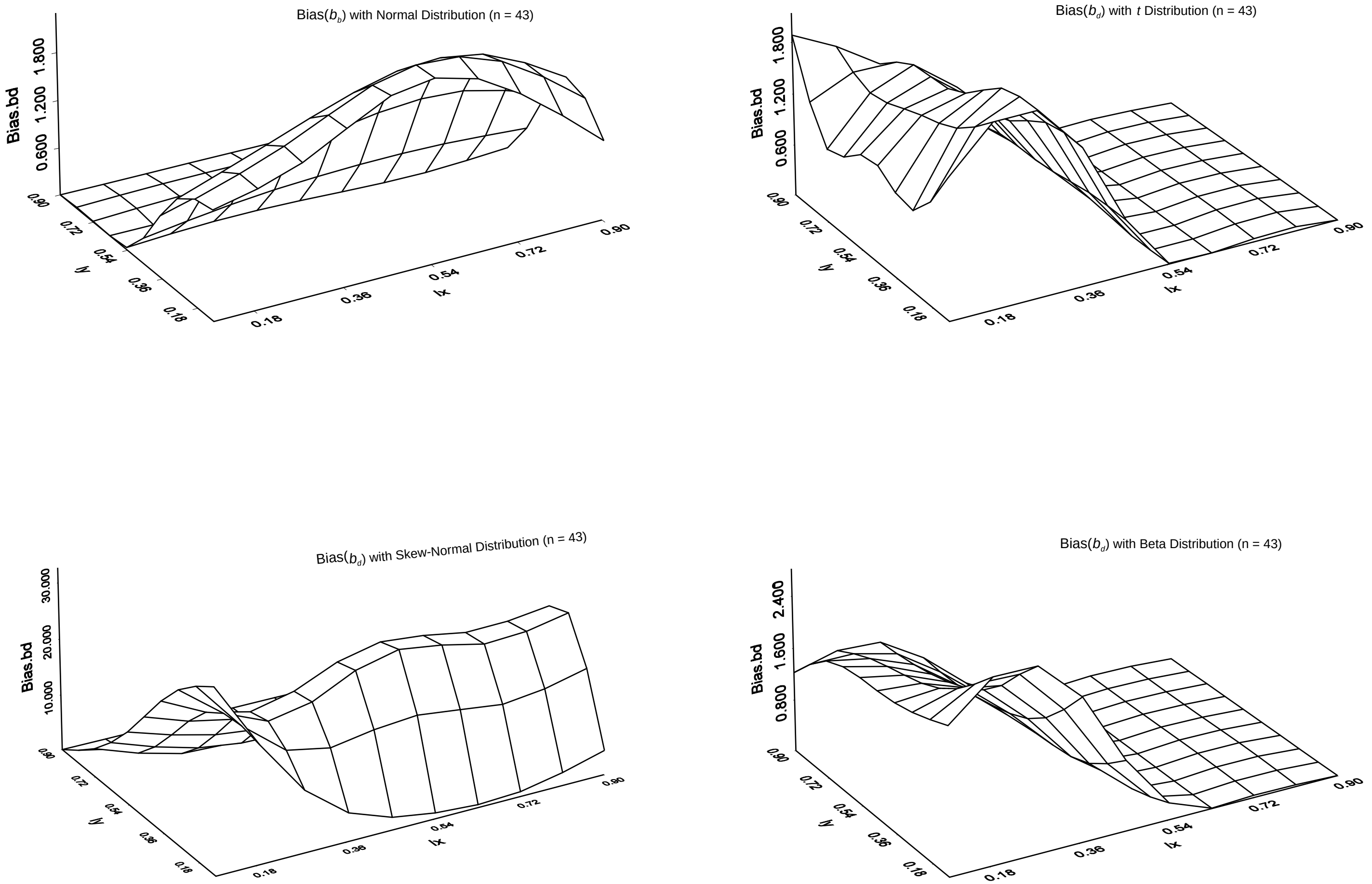
5. Weibull distribution.

Sample sizes $n = 25$ (treated as small sample) and $n = 43$ (treated as large sample) are considered.

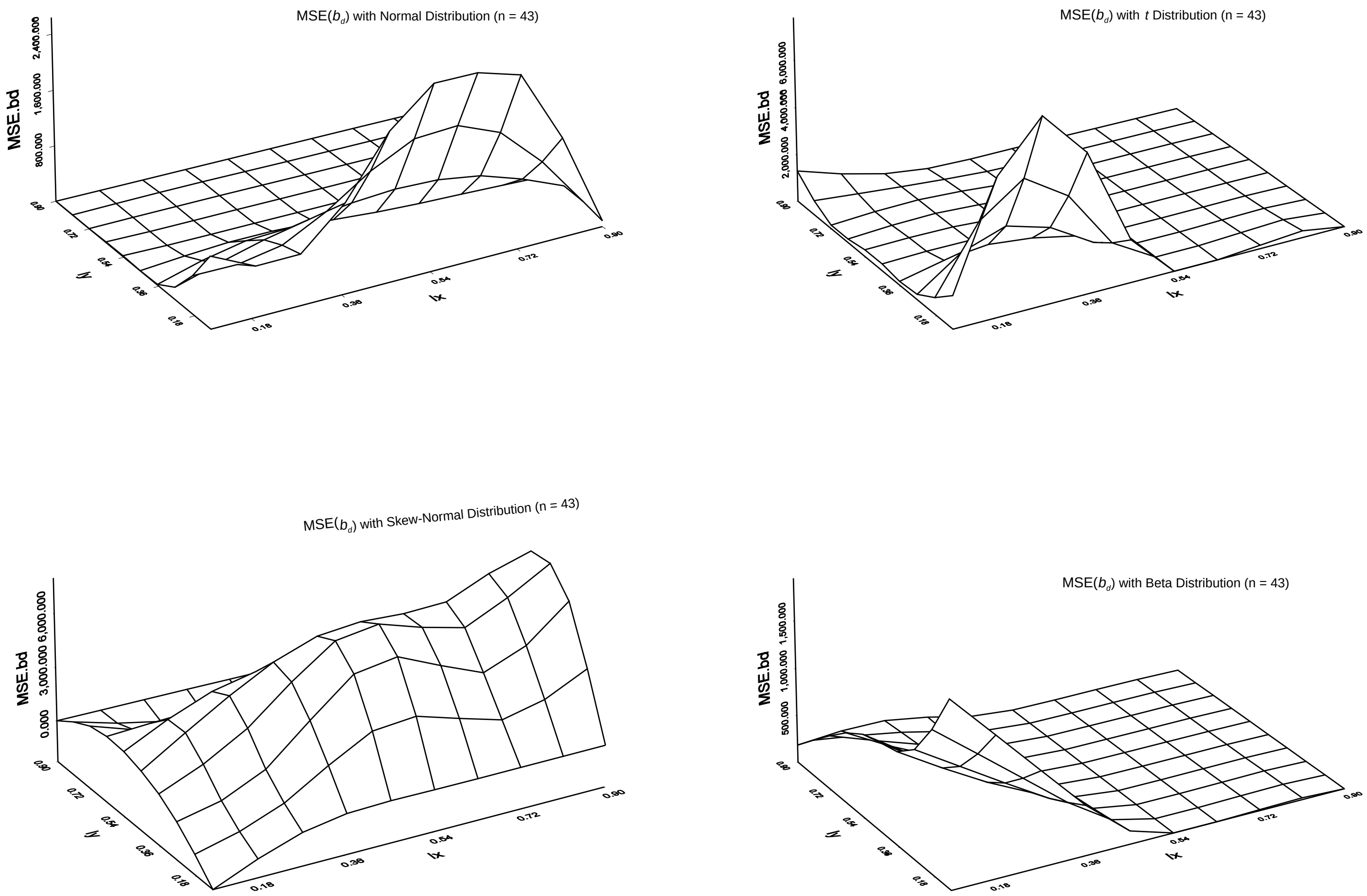
Empirical bias (EB) and empirical mean squared error (EMSE) of b_d, b_i, b_p, b_r, b_m and b_b are computed based on 5000 replications for different combinations of $\lambda_x = 0.1, 0.3, 0.5, 0.7, 0.9$ and $\lambda_y = 0.1, 0.3, 0.5, 0.7, 0.9$ under different distributions of measurement errors.

The values EB and EMSE of these estimators are plotted against λ_x and λ_y in a 3-dimensional surface plot.

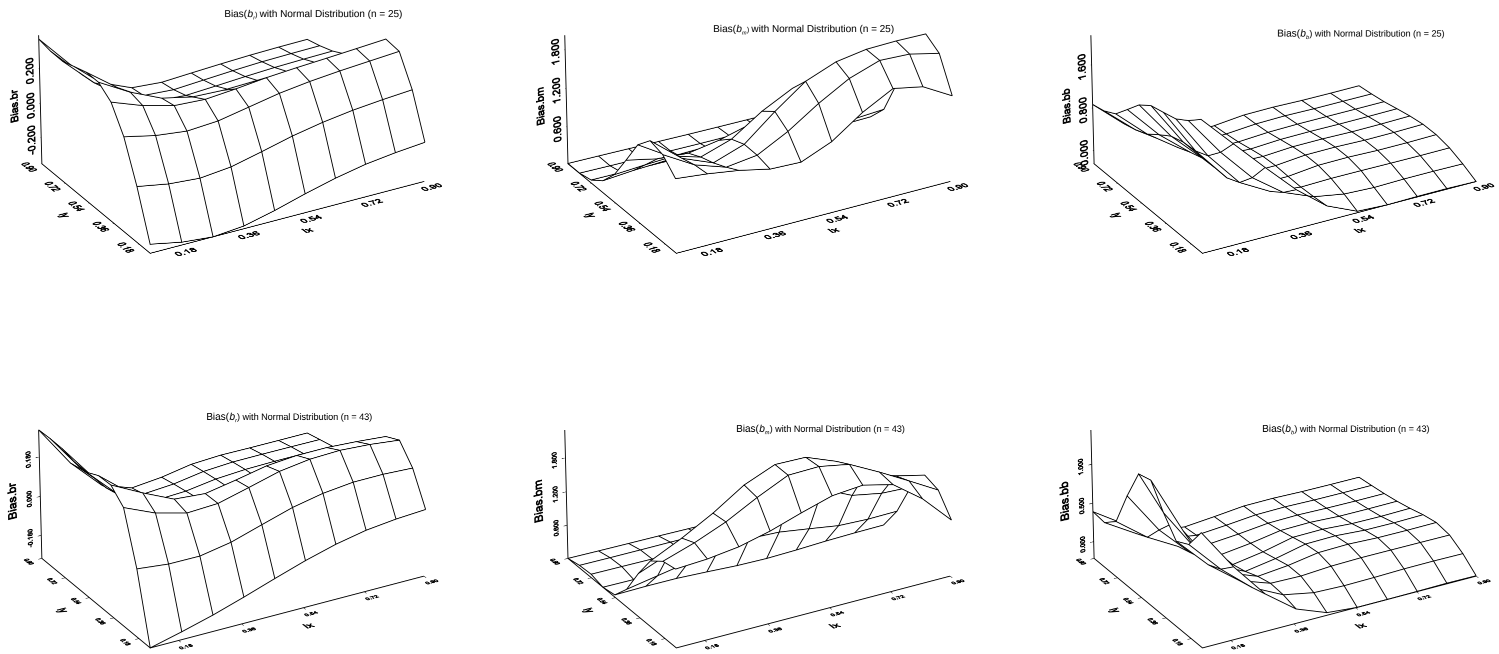
Difference in two surface plots under two different distributions of measurement errors $\Rightarrow$ behaviour of the estimator depends on the distribution of measurement errors.



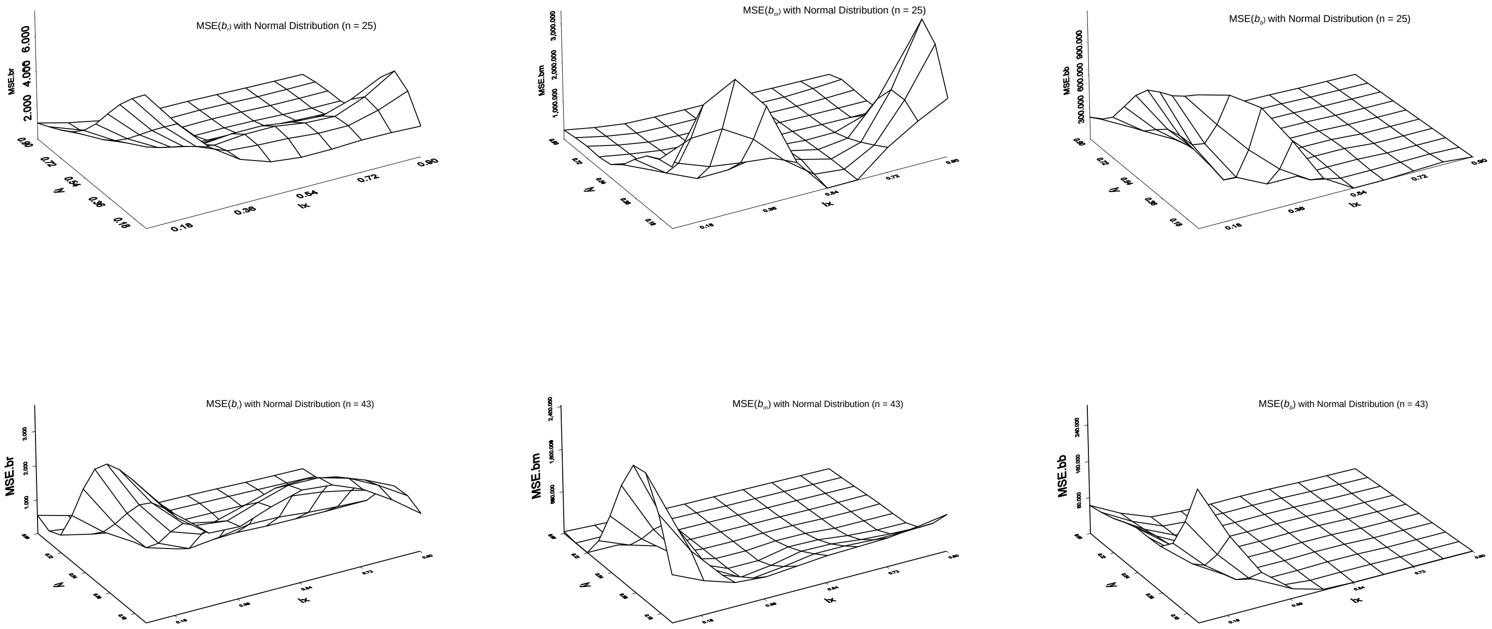
Empirical bias of b_d when measurement errors follow different distributions



Empirical mse of b_d when measurement errors follow different distributions



Empirical bias of estimators when measurement errors follow normal distribution



Empirical mse of estimators when measurement errors follow normal distribution

Empirical bias and empirical mean squared error of estimators under Normal distribution with $n = 25$

λ_x	λ_y	EB(b_d)	EMSE(b_d)	EB(b_i)	EMSE(b_i)	EB(b_p)	EMSE(b_p)	EB(b_r)	E
0.1	0.1	2.089	1172.2173	-1.3759	986.6908	3.4047	2515.2532	-0.3015	
0.1	0.3	1.7396	421.1165	1.2588	6543.8844	7.1188	7082.3023	0.2304	
0.1	0.5	0.6163	49.4849	1.0561	3011.413	15.193	63950.4496	0.3911	
0.1	0.7	0.0841	1.3815	0.5237	2230.9995	23.5769	53226.1105	0.3484	
0.1	0.9	0.0111	0.4415	2.1354	1731.3322	101.4696	2364682.359	0.3692	
0.3	0.1	1.0035	3487.9245	0.0095	12390.1999	3.4205	46690.3197	-0.353	
0.3	0.3	2.2302	1139.6871	-0.049	689.9581	1.5118	933.892	0.167	
0.3	0.5	0.6255	32.431	0.1524	81.473	0.9368	367.1473	0.1686	
0.3	0.7	0.0921	0.3203	0.1575	55.5964	1.03	584.3253	0.0405	
0.3	0.9	0.0161	0.1262	0.0917	1.0878	0.4487	232.9447	0.01	
0.5	0.1	0.9713	260.2454	-0.0198	563.1915	1.3962	504.9342	-0.2695	
0.5	0.3	1.7528	225.9812	0.0442	46.6033	0.4372	36.7253	0.2286	
0.5	0.5	0.7034	81.8598	0.0251	4.3178	0.1257	1.5329	0.1269	
0.5	0.7	0.0864	0.1854	-0.0302	0.2743	0.03	0.3232	-0.0029	
0.5	0.9	0.0137	0.0544	-0.0581	0.1496	0.0035	0.0519	-0.0381	
0.7	0.1	1.6627	975.5491	0.49	1474.0067	1.7782	1106.7196	-0.1717	
0.7	0.3	1.9329	349.1064	0.2131	29.5023	0.3467	23.0407	0.2788	
0.7	0.5	0.5998	39.6491	0.027	0.4387	0.0623	0.1867	0.1425	
0.7	0.7	0.0875	0.1319	-0.0119	0.0793	0.0229	0.056	0.0256	
0.7	0.9	0.017	0.0285	-0.0308	0.0456	0.0061	0.0262	-0.0108	
0.9	0.1	1.3506	417.5147	0.9993	5854.8029	2.1225	5206.1129	-0.1254	
0.9	0.3	1.0239	199.6654	0.0509	103.5534	0.2089	6.0219	0.312	
0.9	0.5	0.6573	115.8942	0.0443	0.0878	0.0534	0.0878	0.1717	
0.9	0.7	0.0891	0.106	0.0114	0.0309	0.0205	0.03	0.0441	
0.9	0.9	0.0155	0.0122	-0.0044	0.0116	0.0049	0.0103	0.0049	

Conclusions:

- The effect of non-normality is present and it affects the performance of estimators. Kurtosis shows more effect than skewness.

- No unique dominance of any estimator. It depend on the choice of distribution of measurement errors and values of reliability ratios.

- For lower values of λ_x and λ_y : More care is needed in choosing the estimator.

Comparison of Normal and t surface plots

EB : Surface plots of all estimators are different

EMSE : Only b_r and b_m have similar surface plots. Others are different

Comparison of Normal and Skew-normal surface plots

EB : b_d, b_p and b_r have similar surface plots whereas b_i, b_b and b_m have different surface plots.

EMSE : b_p has similar surface plots. Rest are different.

Comparison of Normal and Beta surface plots

EB and EMSE : All surface plots are different.

Normal Distribution:

– b_d is more affected by λ_y.

– b_i is more affected by λ_x.

– When λ_x and λ_y are low $\rightarrow$ b_r is better than other estimators.

– When λ_x and λ_y are high $\rightarrow$ Performance of all estimators stablize.

t **Distribution:**

- b_d is more affected by λ_x.
- b_i is more affected by λ_y.
- b_d and b_i are not good for lower values of λ_x and λ_y.
- When $\lambda_x \leq 0.5$ and $\lambda_y \leq 0.5$, b_p and $b_m \rightarrow$ high EB and EMSE.
- When $\lambda_x \geq 0.5$ and $\lambda_y \geq 0.5$, b_r and $b_p \rightarrow$ lower EB and EMSE.

Skew-normal Distribution:

- b_d, b_p and b_m $\rightarrow$ EB and EMSE are high for lower values of λ_x and λ_y.

- b_p is winner for most combinations of λ_x and λ_y.

- When λ_x is very low (say, 0.1) $\rightarrow$ Unstable performance of all estimators.

Beta Distribution:

- When $\lambda_x \leq 0.3 \rightarrow b_d$ has high EMSE.
- When $\lambda_x \leq 0.3$ and $\lambda_y \leq 0.3 \rightarrow b_i$, b_p and b_m have high EMSE.
- When $\lambda_x \leq 0.3$ and $\lambda_y \leq 0.1 \rightarrow b_b$ has high EMSE.
- Overall b_r has smaller EMSE even for lower λ_x and λ_y.

Weibull Distribution:

– Not good conditions.

– Extreme value presence.

– J- shaped curve

References

Aaronsom, M., G. Bothum, J. Mouldt, J. Huchra, R.A. Schommer and M.E. Cornell (1986): 'A distance scale from the infrared magnitude /HI velocity-width relation. ν. distance moduli to 10 galaxy clusters, and positive detection of bulk supercluster motion toward the microwave anisotrophy', *Astrophysical Journal*, **302**, pp. 536-563.

Gleser, L.J. (1992): 'The importance of assessing measurement reliability in multivariate regression', *Journal of American Statistical Association*, **87**, 419, pp. 696-407.

Gleser, L.J. (1993): 'Estimation of slope in linear errors-in-variables regression models when the predictors have known reliability matrix', *Statistics and Probability Letters*, **17 (2)**, pp. 113-121.

Pierce, M.J. and R.B. Tully (1988): 'Distances to virgo and ursa major clusters and a determination of H_0', *Astrophysical Journal*, **330**, pp. 579-595.

Sokal, R.R. and F.J. Rohalf (1981): **Biometry: The principal and Practice of Statistics in Biological Research**, Second Edition, Freeman.