Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 1143 – 1148
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the estimation of the linear relation when the error variances are known

H. Schneeweiss[a], Shalabh[b,*]

[a]Department of Statistics, University of Munich, 80799 Munich, Germany
[b]Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur 208016, India

Available online 28 June 2007

Abstract

The problem of consistent estimation in measurement error models in a linear relation with not necessarily normally distributed measurement errors is considered. Three possible estimators which are constructed as different combinations of the estimators arising from direct and inverse regression are considered. The efficiency properties of these three estimators are derived and the effect of non-normally distributed measurement errors is analyzed. A Monte-Carlo experiment is conducted to study the performance of these estimators in finite samples.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Measurement error models; Ultrastructural model; Non-normal errors

1. Introduction

In a linear measurement error model, the parameters can be estimated consistently only when some additional information besides the data set is available. There are various ways in which such additional information can be employed; (see, e.g., Cheng and Van Ness, 1999; Fuller, 1987). Among them, application of the knowledge of all or one of the measurement error variances is the most prominent approach.

We consider three different combinations of the direct and inverse adjusted least squares (LS) estimators. They are modelled after analogous combinations found in the literature, where, however, they have been constructed from non-adjusted direct and inverse LS estimators. Sokal and Rohalf (1981) considered the geometric mean of these two estimators (which they call the technique of reduced major axis) and Aaronsom et al. (1986) work with the arithmetic mean. In addition, the slope parameter may be 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). While all these estimators are not consistent (although they possibly reduce the bias inherent in their constituent direct and inverse LS estimators), the present paper constructs consistent estimators by using error adjusted direct and inverse LS rather than non-adjusted direct and inverse LS estimators. A simple question then arises: which out of these suggested estimators is better under what conditions. This question has been partly dealt with in Dorff and Gurland (1961), but for a model with replicated observations and unknown error variances.

* Corresponding author. Tel.: +91 512 2597905; fax: +91 512 2597500.
E-mail addresses: shalab@iitk.ac.in, shalabh1@yahoo.com (Shalabh).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.06.018

The efficiency properties of all the estimators under consideration are expressed as functions of the reliability ratios associated with study and explanatory variables, (see, Gleser, 1992, 1993). The asymptotic properties of the estimators are derived when the measurement errors are not necessarily normally distributed.

The plan of our presentation is as follows. In Section 2, we describe a linear model with measurement errors and present the estimators of the slope parameter when the error variances are known. Section 3 analyzes the asymptotic properties of the estimators when the underlying error distributions are not necessarily normal. The finite sample properties of the proposed estimators under different types of distributions of measurement errors are studied through a Monte-Carlo experiment in Section 4. Some concluding remarks are offered in Section 5.

2. Model specification and the estimators

Consider a linear measurement error model in which the variables are related by the linear relation

$$Y_j = \alpha + \beta X_j \quad (j = 1, 2, \ldots, n), \tag{1}$$

where Y_j and X_j denote the true but unobserved values of the study and explanatory variables. The observed values y_j and x_j are expressible as $y_j = Y_j + u_j$ and $x_j = X_j + v_j$, respectively, where u_j and v_j denote the associated measurement errors.

We assume that $X_1, X_2, \ldots, X_n$ are independent (not necessarily identically distributed) random variables such that $\text{plim}_{n\to\infty} \bar{X}$ and $\text{plim}_{n\to\infty}(1/n)\sum(X_j - \bar{X})^2 > 0$ exist, which are denoted by μ_X and σ_X^2, respectively. The measurement errors $u_1, u_2, \ldots, u_n$ are assumed to be independent and identically distributed with mean 0, variance σ_u^2, third moment $\gamma_{1u}\sigma_u^3$ and fourth moment $(\gamma_{2u}+3)\sigma_u^4$. Similarly, the errors $v_1, v_2, \ldots, v_n$ are assumed to be independent and identically distributed with mean 0, variance σ_v^2, third moment $\gamma_{1v}\sigma_v^3$ and fourth moment $(\gamma_{2v}+3)\sigma_v^4$. Further, the random variables (X_j, u_j, v_j) are assumed to be jointly independent.

It may be noted that this model comprises the ultrastructural model, see Dolby (1976), which in turn contains the structural and the functional model as special cases.

Consistent estimation of the parameters α and β in the relationship (1) with the help of given data (x_j, y_j), $j=1, \ldots, n$, is possible only when some additional information is available. This additional information, let us suppose, specifies the error variances σ_u^2 and σ_v^2. We can then estimate the slope parameter β consistently by using the knowledge of either of the two error variances. This provides the following well-known estimators of β:

$$b_d = \frac{s_{xy}}{s_{xx} - \sigma_v^2} \quad \text{and} \quad b_i = \frac{s_{yy} - \sigma_u^2}{s_{xy}},$$

where $s_{xx} = (1/n)\sum(x_j - \bar{x})^2$, $s_{yy} = (1/n)\sum(y_j - \bar{y})^2$, $s_{xy} = (1/n)\sum(x_j - \bar{x})(y_j - \bar{y})$, $\bar{x} = (1/n)\sum x_j$, and $\bar{y} = (1/n)\sum y_j$. An estimator using the knowledge of both the error variances is given by

$$b_p = t_p + \left(t_p^2 + \frac{\sigma_u^2}{\sigma_v^2}\right)^{1/2} \quad t_p = \frac{1}{2s_{xy}}\left(s_{yy} - \frac{\sigma_u^2}{\sigma_v^2} s_{xx}\right). \tag{2}$$

We can combine the two basic estimators b_d and b_i in various ways. One possibility is to estimate the slope parameter β by the geometric mean of the estimators b_d and b_i:

$$b_g = \text{sign}(s_{xy})|b_d b_i|^{1/2}. \tag{3}$$

Similarly, we may estimate β by the arithmetic mean of b_d and b_i:

$$b_m = \tfrac{1}{2}(b_d + b_i). \tag{4}$$

Another interesting estimator of β is

$$b_b = t_b + (t_b^2 + 1)^{1/2} \quad t_b = \frac{b_d b_i - 1}{b_d + b_i}, \tag{5}$$

which is the slope of the line that bisects the angle between the two regression lines specified by b_d and b_i.

3. Asymptotic properties

The asymptotic variances of the basic estimator b_d, b_i and b_p under an ultrastructural model and when errors are not necessarily normally distributed have been studied by Shalabh et al. (2004), (see also Srivastava and Shalabh, 1997; Schneeweiss, 1976; Fuller, 1987). For later reference, we restate these results. In addition, we give an expression for the asymptotic covariance of b_d and b_i, which will be used in the derivation of the asymptotic variances of b_g, b_m and b_b.

Proposition 1. *The estimators b_d and b_i are asymptotically jointly normally distributed as*

$$\sqrt{n}\begin{pmatrix} b_d-\beta \\ b_i-\beta \end{pmatrix} \to \mathrm{N}(0,\Sigma_b) \quad \textit{where } \Sigma_b=\begin{pmatrix} \sigma_{dd} & \sigma_{di} \\ \sigma_{di} & \sigma_{ii} \end{pmatrix}$$

with

$$\sigma_{dd}=\beta^2\left(\frac{1-\lambda_x}{\lambda_x^2}\right)[\lambda_x+q+(1-\lambda_x)(2+\gamma_{2v})], \tag{6}$$

$$\sigma_{ii}=\beta^2\left(\frac{1-\lambda_x}{\lambda_x^2}\right)[\lambda_x+q+q^2(1-\lambda_x)(2+\gamma_{2u})] \tag{7}$$

and

$$\sigma_{di}=\beta^2\left(\frac{1-\lambda_x}{\lambda_x^2}\right)[\lambda_x+q(2\lambda_x-1)], \tag{8}$$

where $\lambda_x=\sigma_X^2/(\sigma_X^2+\sigma_v^2)$, $\lambda_y=\beta^2\sigma_X^2/(\beta^2\sigma_X^2+\sigma_u^2)$ *and* $q=\sigma_u^2/(\beta^2\sigma_v^2)=\lambda_x(1-\lambda_y)/\lambda_y(1-\lambda_x)$.

Proof. For σ_{dd} and σ_{ii}, see Shalabh et al. (2004). The covariance σ_{di} can be derived in a similar way (see also Schneeweiss and Shalabh, 2006).

Notice that λ_x and λ_y are the reliability ratios of the explanatory and study variables in the model. Obviously, $0<\lambda_x\leqslant 1$, $0\leqslant\lambda_y\leqslant 1$, and $q\geqslant 0$.

Similarly, $\sqrt{n}(b_p-\beta)\to \mathrm{N}(0,\sigma_{pp})$, where the asymptotic variance is given by

$$\sigma_{pp}=\beta^2\left(\frac{1-\lambda_x}{\lambda_x^2}\right)\left[\lambda_x+q+\frac{q^2(1-\lambda_x)}{(q+1)^2}(\gamma_{2u}+\gamma_{2v})\right]. \quad \square \tag{9}$$

We need the following general result to derive the asymptotic variance of the remaining estimators.

Proposition 2. *Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be two consistent, asymptotically jointly normal estimators of β. Any estimator $\hat{\beta}$ of β which is a differentiable and symmetric function $g(\hat{\beta}_1,\hat{\beta}_2)$ of $\hat{\beta}_1$ and $\hat{\beta}_2$ such that $\beta=g(\beta,\beta)$ is consistent and asymptotically normally distributed with an asymptotic variance given by*

$$\sigma_{\hat{\beta}}^2=\tfrac{1}{4}(\sigma_{11}+2\sigma_{12}+\sigma_{22}),$$

where $\Sigma=(\sigma_{ij})$, $i,j=1,2$, is the asymptotic covariance matrix of $(\hat{\beta}_1,\hat{\beta}_2)$.

Proof. Denote the partial derivatives of g with respect to the first and second argument of g by g_1 and g_2, respectively. Then by the symmetry of g, the equation $\beta=g(\beta,\beta)$ implies

$$1=g_1(\beta,\beta)+g_2(\beta,\beta)=2g_1(\beta,\beta)=2g_2(\beta,\beta),$$

i.e., $g_1(\beta, \beta) = g_2(\beta, \beta) = \frac{1}{2}$. We then can evaluate $\sigma^2_{\hat{\beta}}$ by the delta-method as

$$\sigma^2_{\hat{\beta}} = \tfrac{1}{4}(1, 1)\Sigma(1, 1)^\top$$

which is the statement of Proposition 2. □

As a consequence of Propositions 1 and 2, the estimators b_g, b_m, and b_b are all consistent and asymptotically normal with the same asymptotic variance given by

$$\beta^2 \frac{(1-\lambda_x)\delta}{4\lambda_x^2}, \tag{10}$$

where

$$\delta = 2[q^2(1-\lambda_x) + 1 + \lambda_x + 2q\lambda_x] + (1-\lambda_x)(\gamma_{2v} + q^2\gamma_{2u}). \tag{11}$$

It is interesting to observe from (10) and (11) that only the kurtosis of the distributions of measurement errors shows its influence on the asymptotic variances of these estimators.

Comparing the asymptotic variances, we find that the estimator b_p is more efficient than the equally efficient estimators b_g, b_m or b_b if, and only if,

$$2(q^2-1)^2 \geqslant (q-1)[(1+3q)\gamma_{2v} - q^2(q+3)\gamma_{2u}]. \tag{12}$$

Condition (12) is clearly satisfied when both measurement errors have mesokurtic (e.g., normal) distributions. The condition also holds true when $q > 1$ and $\gamma_{2u} \geqslant 0$ and $\gamma_{2v} < 0$. However in other cases, the estimator b_p may be less efficient under non-normal measurement errors.

Next, we compare the asymptotic variances of b_g, b_m or b_b to b_d and b_i. We find that b_g, b_m or b_b are better than b_d if, and only if,

$$3\gamma_{2v} - q^2\gamma_{2u} > 2(q-3)(q+1). \tag{13}$$

Condition (13) is always satisfied for mesokurtic (e.g., normal) distributions of u and v when $q < 3$.

Similarly, b_g, b_m or b_b are better than b_i if, and only if,

$$\gamma_{2v} - 3q^2\gamma_{2u} < 2(3q-1)(q+1). \tag{14}$$

Condition (14) is always satisfied for mesokurtic (e.g., normal) distributions of u and v when $q > \frac{1}{3}$.

Although the estimators b_g, b_m and b_b use both type of additional information on measurement error variances, they are sometimes less efficient than b_d and b_i, which use information only on one measurement error variance.

4. Monte-Carlo simulation

We conducted a Monte-Carlo simulation to study the behavior of the estimators in finite samples. The following probability distributions of measurement errors are considered to give an idea of the effect of departure from the normal distribution on the efficiency properties of the estimators: normal distribution (having no skewness and no kurtosis) and t-distribution with six degrees of freedom (having no skewness but non-zero kurtosis).

Two data sets of sample sizes $n = 40$ (treated as small sample) and $n = 400$ (treated as large sample) are considered. The model parameters are $\alpha = 1$, $\beta = 2$, $\mu_X = 10$, and $\sigma_X^2 = 0.08$. The empirical bias (EB) and empirical mean squared error (EMSE) of the estimators b_d, b_i, b_p, b_g, b_m, and b_b were computed based on 10 000 replications for both the sample sizes and for different combinations of λ_x and λ_y under different distributions of measurement errors. Some selected values of EB and EMSE of these estimators for $n = 40$ and $n = 400$ are presented in Tables 1 and 2 for normal and t-distributed measurement errors. We also plotted the EB and EMSE of all the estimators against λ_x and λ_y in three-dimensional surface plots. In order to shorten the discussion and save space, we are not presenting them here. They can be found in an extended version of this section in Schneeweiss and Shalabh (2006) where further discussions including an analysis of the beta distribution $Beta(4, 2)$ and the Weibull distribution $W(1, 2)$ can also be found.

Table 1
Empirical bias and empirical mean squared error of estimators under normal distribution with $n = 40$ and 400

λ_x	λ_y	EB(b_d)	EB(b_i)	EB(b_p)	EB(b_g)	EB(b_m)	EB(b_b)	EM(b_d)	EM(b_i)	EM(b_p)	EM(b_g)	EM(b_m)	EM(b_b)
When $n = 40$													
0.3	0.3	1.2964	0.1434	0.1897	0.1782	0.7199	0.1098	551.5361	18.9549	4.5883	1.5134	141.6754	0.3282
0.3	0.5	1.2066	0.0757	0.1019	0.2178	0.6412	0.1270	193.5423	7.6577	4.4940	1.1156	49.3896	0.2566
0.3	0.7	1.7229	0.0462	0.0541	0.2679	0.8846	0.1423	539.3231	0.3843	0.2092	1.6768	132.9626	0.2586
0.5	0.3	0.1369	0.0160	0.0648	0.0146	0.0765	0.0199	0.2907	9.8367	0.7767	0.1461	2.5301	0.1241
0.5	0.5	0.2153	−0.0013	0.0337	0.0501	0.1070	0.0410	10.2123	0.1400	0.0868	0.1738	2.5832	0.0995
0.5	0.7	1.0319	0.0385	0.0557	0.1801	0.5352	0.1161	495.7989	0.3371	0.2566	1.0050	123.4030	0.1862
0.7	0.3	0.0872	0.1391	0.1757	0.0130	0.1132	0.0234	0.1468	41.0841	34.1450	0.1577	10.2917	0.1304
0.7	0.5	0.0516	−0.0271	0.0187	−0.0108	0.0123	−0.0077	0.0841	0.1464	0.0598	0.0744	0.0692	0.0688
0.7	0.7	0.0401	−0.0081	0.0094	0.0106	0.0160	0.0104	0.0481	0.0384	0.0290	0.0310	0.0303	0.0306
When $n = 400$													
0.3	0.3	0.0648	0.0081	0.0151	0.0168	0.0365	0.0158	0.1098	0.0713	0.0335	0.0382	0.0847	0.0357
0.3	0.5	0.0566	0.0071	0.0100	0.0221	0.0319	0.0211	0.0788	0.0226	0.0166	0.0236	0.0274	0.0223
0.3	0.7	0.0575	0.0064	0.0082	0.0254	0.0320	0.0246	0.0659	0.0107	0.0100	0.0198	0.0225	0.0189
0.5	0.3	0.0115	−0.0033	0.0033	−0.0042	0.0041	−0.0039	0.0216	0.0489	0.0156	0.0190	0.0189	0.0186
0.5	0.5	0.0125	−0.0001	0.0031	0.0030	0.0062	0.0030	0.0158	0.0155	0.0089	0.0090	0.0092	0.0090
0.5	0.7	0.0160	0.0015	0.0036	0.0071	0.0088	0.0071	0.0125	0.0061	0.0052	0.0060	0.0061	0.0060
0.7	0.3	0.0030	−0.0095	0.0007	−0.0086	−0.0033	−0.0083	0.0103	0.0410	0.0095	0.0158	0.0153	0.0154
0.7	0.5	0.0028	−0.0038	0.0005	−0.0016	−0.0005	−0.0016	0.0049	0.0088	0.0042	0.0047	0.0047	0.0047
0.7	0.7	0.0043	−0.0009	0.0013	0.0013	0.0017	0.0013	0.0031	0.0030	0.0023	0.0023	0.0023	0.0023

Table 2
Empirical bias and empirical mean squared error of estimators under t-distribution with $n = 40$ and 400

λ_x	λ_y	EB(b_d)	EB(b_i)	EB(b_p)	EB(b_g)	EB(b_m)	EB(b_b)	EM(b_d)	EM(b_i)	EM(b_p)	EM(b_g)	EM(b_m)	EM(b_b)
When $n = 40$													
0.3	0.3	1.6511	0.0865	0.2307	0.1731	0.8688	0.1116	4896.7476	978.6324	15.8123	1.3589	1466.8572	0.3056
0.3	0.5	1.8224	0.2355	0.3630	0.1655	1.0290	0.0788	521.2361	602.0913	69.3003	1.8048	278.6083	0.3780
0.3	0.7	1.4976	0.0403	0.0488	0.2539	0.7690	0.1465	625.3957	0.1906	0.1417	1.5268	154.9934	0.2502
0.5	0.3	0.1755	0.0857	0.0857	0.0250	0.1306	0.0290	0.5406	2.9772	1.2748	0.1888	0.8544	0.1528
0.5	0.5	0.3019	0.0049	0.0422	0.0572	0.1534	0.0448	42.2078	0.2295	0.1146	0.2814	10.5862	0.1255
0.5	0.7	0.3194	0.0192	0.0404	0.1040	0.1693	0.0837	5.2565	0.0884	0.0728	0.2352	1.3152	0.1202
0.7	0.3	0.0783	0.0425	0.0855	−0.0064	0.0604	0.0037	0.1262	1.2832	6.1390	0.1402	0.3532	0.1185
0.7	0.5	0.0467	−0.0340	0.0141	−0.0119	0.0064	−0.0099	0.0720	0.1214	0.0521	0.0637	0.0598	0.0602
0.7	0.7	0.0375	−0.0062	0.0108	0.0117	0.0157	0.0116	0.0375	0.0308	0.0239	0.0253	0.0246	0.0250
When $n = 400$													
0.3	0.3	0.0515	0.0072	0.0116	0.0123	0.0294	0.0118	0.0814	0.0630	0.0292	0.0318	0.0350	0.0304
0.3	0.5	0.0622	0.0092	0.0131	0.0260	0.0357	0.0250	0.0775	0.0232	0.0172	0.0243	0.0275	0.0232
0.3	0.7	0.0512	0.0081	0.0098	0.0241	0.0297	0.0235	0.0533	0.0102	0.0094	0.0171	0.0189	0.0165
0.5	0.3	0.0155	−0.0010	0.0067	−0.0003	0.0073	−0.0001	0.0210	0.0438	0.0147	0.0175	0.0173	0.0171
0.5	0.5	0.0126	0.0009	0.0042	0.0043	0.0068	0.0043	0.0128	0.0117	0.0072	0.0073	0.0074	0.0073
0.5	0.7	0.0127	0.0012	0.0031	0.0058	0.0070	0.0058	0.0098	0.0052	0.0045	0.0050	0.0051	0.0050
0.7	0.3	0.0047	−0.0079	0.0022	−0.0058	−0.0016	−0.0056	0.0090	0.0316	0.0082	0.0125	0.0122	0.0123
0.7	0.5	0.0028	−0.0026	0.0007	−0.0012	0.0001	−0.0011	0.0053	0.0097	0.0045	0.0050	0.0050	0.0050
0.7	0.7	0.0035	−0.0004	0.0011	0.0011	0.0016	0.0011	0.0036	0.0035	0.0026	0.0026	0.0026	0.0026

For large samples ($n = 400$), most of the estimators show a behavior that corresponds to the asymptotic theory, at least for higher values of reliability ratios ($\lambda \geqslant 0.5$). In particular, the MSEs of b_g, b_m, and b_b are more or less the same and the MSE of b_p is smallest, not only for the normal but also for the t-distribution of the measurement errors. There are slight differences in the variances of the estimators when one goes from the normal error distribution to the *t*-distribution. These differences are due to the peakedness of the *t*-distribution. In most cases, the estimates have a negligible bias. Only for small $\lambda(\lambda < 0.5)$, the bias sometimes becomes noticeable, but even then it is rather small.

For small samples ($n = 40$), results can differ considerably from those of asymptotic theory, in particular if λ_x and/or λ_y are small. We then find high values for the bias and in some cases huge MSE values. Also the uniform superiority of b_p is questionable when the λ's are small.

5. Conclusion

We considered six consistent and asymptotically normally distributed estimators for the slope parameter β when the error variances are known in a linear ultrastructural model.

When the distributions of errors depart from normality, we observe that the asymptotic variances of the estimators are influenced by the peakedness, and not the asymmetry, of the error distributions. Further it is seen that the superiority of an estimator over another under the popular specification of normality may not necessarily carry over when the distributions depart from normality.

These theoretical results are born out by a Monte-Carlo simulation study, although for small samples, the results may deviate from the asymptotic ones, both with regard to the bias and to the MSE.

Acknowledgment

The second author acknowledges the support from a SERC project from the Department of Science and Technology, Government of India, SFB 386, University of Munich (Germany), and AvH Foundation (Germany).

References

Aaronsom, M., Bothum, G., Moultd, J., Huchra, J., Schommer, R.A., Cornell, M.E., 1986. A distance scale from the infrared magnitude/HI velocity-width relation, *v* distance moduli to 10 galaxy clusters, and positive detection of bulk supercluster motion toward the microwave anisotrophy. Astrophys. J. 302, 536–563.

Cheng, C.-L., Van Ness, J.W., 1999. Statistical Regression with Measurement Error. Arnold, London.

Dolby, G.R., 1976. The ultrastructural relation: a synthesis of the functional and structural relations. Biometrika 63, 39–50.

Dorff, M., Gurland, J., 1961. Estimation of parameters of a linear functional relation. J. Roy. Statist. Soc. Ser. B 23 (1), 160–170.

Fuller, W.A., 1987. Measurement Error Models. Wiley, New York.

Gleser, L.J., 1992. The importance of assessing measurement reliability in multivariate regression. J. Amer. Statist. Assoc. 87 (419), 696–707.

Gleser, L.J., 1993. Estimation of slope in linear errors-in-variables regression models when the predictors have known reliability matrix. Statist. Probab. Lett. 17 (2), 113–121.

Pierce, M.J., Tully, R.B., 1988. Distances to virgo and ursa major clusters and a determination of H_0. Astrophys. J. 330, 579–595.

Schneeweiss, H., 1976. Consistent estimation of a regression with errors in the variables. Metrika 23, 101–115.

Schneeweiss, H., Shalabh, 2006. On the Estimation of the Linear Relation When the Error Variances are Known. Universität München, München, Germany (⟨www.stat.uni-muenchen.de/sfb386⟩), Discussion Paper 493.

Shalabh, Gleser, L.J., Rosen, O., 2004. On the usefulness of knowledge of error variances in the consistent estimation of an unreplicated ultrastructural model. J. Statist. Comput. Simulation 74 (6), 391–417.

Sokal, R.R., Rohalf, F.J., 1981. Biometry: The Principal and Practice of Statistics in Biological Research. second ed.. Freeman, New York.

Srivastava, A.K., Shalabh, L., 1997. Consistent estimation for the non-normal ultrastructural model. Statist. Probab. Lett. 34, 67–73.