International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:8, No:8, 2014

Orthogonal Regression for Nonparametric Estimation of Errors-in-Variables Models

Anastasiia Yu. Timofeeva

***Abstract*—**Two new algorithms for nonparametric estimation of errors-in-variables models are proposed. The first algorithm is based on penalized regression spline. The spline is represented as a piecewise-linear function and for each linear portion orthogonal regression is estimated. This algorithm is iterative. The second algorithm involves locally weighted regression estimation. When the independent variable is measured with error such estimation is a complex nonlinear optimization problem. The simulation results have shown the advantage of the second algorithm under the assumption that true smoothing parameters values are known. Nevertheless the use of some indexes of fit to smoothing parameters selection gives the similar results and has an oversmoothing effect.

***Keywords*—**Grade point average, orthogonal regression, penalized regression spline, locally weighted regression.

I. Introduction

THE selection in higher education is based on the assumption that the criterion for students' selecting (unified state exam score) correlates with their performance (first-year university grade point average – FYGPA). The relationship between these parameters is usually assumed to be linear and significant [1]. Only a few work [2], [3] has been devoted to examining of nonlinearity of this relationship. This results appear inconclusively because it is ignored the input factor measurement error.

Currently methods of nonparametric estimation error-in-variables models have been actively developing [4]. This development is aimed at integration with known methods of solving the problem of recovering the structural dependency. Now there are approaches which require extensive additional information (instrumental variables, repeated observations [5]) that leads to the costs of gathering such data. This paper is focused on the total least squares method [6]. It requires only fix the value of the ratio of errors variances, which can be set on the basis of a priori notions of the researcher.

II. Problem Definition

The subject of interest is the shape of the relationship between university grade point average Y and unified state exam score X. Let the unknown functional dependence $Y = g(X)$ be postulated. The observed values x and y of the variables are recorded with random errors ε_x and ε_y with zero expectation:

$$x = X + \varepsilon_x,\ y = Y + \varepsilon_y. \tag{1}$$

The result is a structural equation of the form

$$y = g(x - \varepsilon_x) + \varepsilon_y. \tag{2}$$

The problem is to estimate the response values well consistent with the true values Y. It is well known that the use of the ordinary least squares for identification of the structural equation (2) results in biased and inconsistent estimates [7]. Therefore, to estimate so called errors-in-variables model a number of special approaches are used [8], [9].

III. Estimation Methods

For estimating errors-in-variables model a priori information about the random errors distribution is required. It is assumed that the ratio of the error variances is a given value

$$\gamma = \sigma_{\varepsilon_x}^2 \Big/ \sigma_{\varepsilon_y}^2. \tag{3}$$

Then the estimation problem can be solved by the total least squares method by optimizing the loss function

$$G = \sum_{i=1}^{n} \frac{1}{\gamma} w_i (x_i - X_i)^2 + w_i (y_i - g(X_i))^2 \tag{4}$$

where index i corresponds to the number of observation, n is sample size, w_i is a weight of i th observation point.

The best studied is the case of a linear dependence between two variables, for which there exists an analytical solution [10]. With $\gamma = 1$ such a regression is called orthogonal.

The estimation of nonlinear errors-in-variables models is extremely difficult. There are a number of approaches which combine advantages of orthogonal and nonparametric regression. The author suggests her own estimation algorithms within these approaches.

A. Orthogonal Penalized Regression Spline

One of the widely used nonparametric methods is penalized regression spline (P-spline). In the simple case of a first-order spline the model specifies that for some coefficients vector $\theta = (\theta_0, \theta_1, \ldots, \theta_{K+1})$

A. Yu. Timofeeva is with the Faculty of Business, Novosibirsk State Technical University, 20 Prospekt K. Marksa, Novosibirsk, 630073, Russia (e-mail: a.timofeeva@corp.nstu.ru).

The report study was supported by the Russian Foundation for Basic Research, research project no. 14-07-31171 мол-а.

$$Y = P(X)\theta \tag{5}$$

where $P(X)$ is the spline basis, K is the number of knots. In particular, in [4] a basis is proposed of the form

$$P(X) = \left(1, X, (X-k_1)_+, \ldots, (X-k_K)_+\right) \tag{6}$$

where $k_1 < \ldots < k_K$ are the knots in region of X, z_+ is the positive part of z. The knots are typically determined using sample quantiles of input variable, in particular in [4] k_l is defined as $(l+1)/(K+2)$-th sample quantile of x.

An estimate of the vector θ can be obtained by the least squares method:

$$\hat{\theta} = \left(P(X)'P(X) + \lambda D\right)^{-1} P(X)'y \tag{7}$$

where λ is the smoothing parameter, D is the diagonal matrix with the first two zeroes and K ones on the diagonal called the penalty matrix.

For the P-spline estimation of errors-in-variables model Iterative Conditional Modes algorithm was proposed in [8]. The initial values for the vector θ is estimated by naive P-sline assuming $X = x$. For fixed values of X the optimization problem with loss function (4) is solved by ordinary least squares. So it turns out the next approximation of estimates vector $\hat{\theta}$. Further, for fixed values of the parameters the true values of input variable are estimated by minimizing the function (4). This steps iterate to convergence in estimates vector $\hat{\theta}$. Within the overall algorithm the weight w_i equals to one for all i.

The most difficult part of the algorithm is the estimation at each step of the true values of the input variable. In order to simplify this part, the author suggested in [11] the following procedure. First, the spline is represented as a piecewise-linear function whose parameters $\{\alpha_l, \beta_l\}_{l=\overline{0,K}}$ are calculated from the recurrence relations

$$\alpha_l = \alpha_{l-1} - \theta_{l+1}k_l, \ \beta_l = \beta_{l-1} + \theta_{l+1} \tag{8}$$

for $\alpha_0 = \theta_0$, $\beta_0 = \theta_1$. Then, there are calculated the elements of the matrices of weighted distances from the observed values of the variables to each knot Rk and to each linear portion Rl of the regression curve:

$$Rl_{il} = \begin{cases} \dfrac{|y_i - \alpha_l - \beta_l x_i|}{\sqrt{1+\gamma\beta_l^2}}, & if\ (x_i, y_i) \in U_l, \\ M, & otherwise, \end{cases} \tag{9}$$

$$Rk_{ij} = \sqrt{\frac{1}{\gamma}(x_i - k_j)^2 + (y_i - \alpha_j - \beta_j k_j)^2}, \tag{10}$$

$i = \overline{1,n}$, $l = \overline{0,K}$, $j = \overline{1,K}$, M is the maximally large number, U_l is the set of pairs of the observed values of variables that satisfy the inequality

$$v_l(k_l) < sign(\beta_l)\cdot y_i < v_l\left(k_{l+1}\right), \tag{11}$$

where $v_l(k_j) = sign(\beta_l)\cdot\left(\alpha_l + \beta_l k_j - \frac{1}{\gamma\beta_l}(x_i - k_j)\right)$, $k_0 = -M$, $k_{K+1} = M$.

On the basis of the composite matrix of distances $\tilde{R} = [Rl \mid Rk] = \{\tilde{R}_{is}\}$ of dimensions $N \times (2K+1)$, the number l_i^* is determined of the linear portion of the curve with the minimal distance to the i th observation

$$l_i^* = Arg \min_{s=0,2K} \tilde{R}_{is}, \ s = \overline{0,2K}. \tag{12}$$

Finally, the estimates of the true values of input variable are found according to the following rule:

$$\hat{X}_i = \begin{cases} \dfrac{x_i + \gamma\beta_{l_i^*}(y_i - \alpha_{l_i^*})}{1+\gamma\beta_{l_i^*}^2}, & l_i^* \le K, \\ k_{l_i^*-K}, & l_i^* > K. \end{cases} \tag{13}$$

This algorithm is investigated and applied to the problem Engel curve estimation in [11].

B. Orthogonal Locally Weighted Regression

A somewhat different approach to nonparametric estimation uses locally weighted regression (LOESS). LOESS was proposed in [12]. Its main idea is to construct response estimates at given points $\tilde{X}_j$ of the input variable on the k nearest neighbors. Based on the distance h_j from the j th point to the k th nearest point the weights are calculated

$$w_i\left(\tilde{X}_j\right) = W\left(h_j^{-1}\rho_{\tilde{X}_j X_i}\right) \tag{14}$$

where $\rho_{\tilde{X}_j X_i}$ is Euclidean distance from $\tilde{X}_j$ to i th value of input variable, $j = \overline{1,m}$, $i = \overline{1,n}$. As a weight function $W(z)$ is traditionally used the tricube function

$$W(z) = \left(1 - z^3\right)_+^3. \tag{15}$$

However, any other weight function that satisfies the certain properties could be used, e.g. the rectangular weight function

$$W(z) = H(1-z) \tag{16}$$

where $H(\cdot)$ is the Heaviside function.

Thus, the calculated weights $\{w_i(\tilde{X}_j)\}_{i=\overline{1,n}}$ are used to regression estimate by weighted least squares method. The regression is estimated for each $\tilde{X}_j$.

For errors-in-variables model estimation it seems logical that least squares problem is replaced by minimizing (4) with weight $w_i(\tilde{X}_j)$. In the simplest case local linear approximation with $\gamma = 1$ only it is necessary to estimate the weighted orthogonal regression for each $\tilde{X}_j$.

But the problem lies in the fact that the weights depend on the true values of input variable. To solve this problem the adaptive estimation approach is proposed in [9]. In contrast to this approach, a complex nonlinear optimization problem is considered with the objective function

$$G_j = \frac{1}{1+\gamma\beta_j^2}\sum_{i=1}^{n} w_i(\tilde{X}_j, X_{ij})\left(y_i - \alpha_j - \beta_j x_i\right)^2 \qquad (17)$$

where the ith weight depends on the true values of input variable which is determined by the ratio

$$X_{ij} = \frac{x_i + \gamma\beta_j\left(y_i - \alpha_j\right)}{1+\gamma\beta_j^2}. \qquad (18)$$

Optimization of loss function (17) was carried out using a combination of golden section search and successive parabolic interpolation. The proposed algorithms for errors-in-variables model estimation were implemented using a free software environment for statistical computing R [13].

IV. Smoothing Parameter Selection

An important problem of using of nonparametric estimation methods is the smoothing parameter selection. Standard selection criteria are cross-validation and information criterion [14]. Its calculation is convenient for the linear models when predictions $\hat{y}$ are defined through the hat matrix H as follows $\hat{y} = Hy$.

Unfortunately that is impossible to do for errors-in-variables models. The usual indices of fit can be used, for example $RMSE$, MAE, $MAPE$. Unlike ordinary regression models when the input variable is measured with error the use of these indexes should not lead to the problem of overfitting. This is due to the instability of orthogonal regression estimates for small samples. So if the correlation between the input and the response variables is close to zero, the usual regression line is almost horizontal, whereas the orthogonal regression coefficient will tend to infinity. Therefore orthogonal regression line can greatly deviate from the observation points. The usual indexes of fit should prevent strong deviation.

Also in [8] it is suggested to fit the smoothing parameter using cross-validation. Its value is calculated for initial parameters estimate and is held fixed for the iterative procedure. In particular, leave-one-out cross-validation $LOOCV$ can be used.

V. Data Simulation

Assume that the true dependence of the normalized FYGPA Y on the normalized exam scores X can be described by a model curve

$$Y = X^{\gamma} \qquad (19)$$

where $X \sim B(\alpha, \beta)$, α and β are standard beta distribution parameters, γ indicates differences in a discriminating power of the scores. If $\gamma > 1$ then FYGPA has the greater discriminating power compared to unified state exam scores for the strong students (for large values of X) and the lesser one for the weak (for small values of X), and vice versa.

The true values of both variables are unobservable. The results of observations are random variables (1). Unfortunately the standard assumption of normal error distribution is unacceptable here in so far as observed values must be in the interval [0, 1]. Therefore beta distribution was chosen wherein the parameters have to be dependent on the scores:

$$\varepsilon_x = \varepsilon_x^b - \frac{1}{2},\ \varepsilon_y = \varepsilon_y^b - \frac{1}{2} \qquad (20)$$

where $\varepsilon_x^b \sim B(\alpha_X, \beta_X)$, $\alpha_X = \beta_X = \tau_0 + \tau_1\left|X - E(X)\right|$, and similarly for Y. Parameters τ_0, τ_1 determine the variance of the errors. The parameter values are chosen so that the simulated data are best fit to real: $\gamma = 0.456$, $\alpha = 2.954$, $\beta = 3.48$, $\tau_0 = 3$, $\tau_1 = 50$. Fig. 1 shows that the model and the real scatter plot are very similar. Further 500 samples were generated; volume of each sample was 500 items.

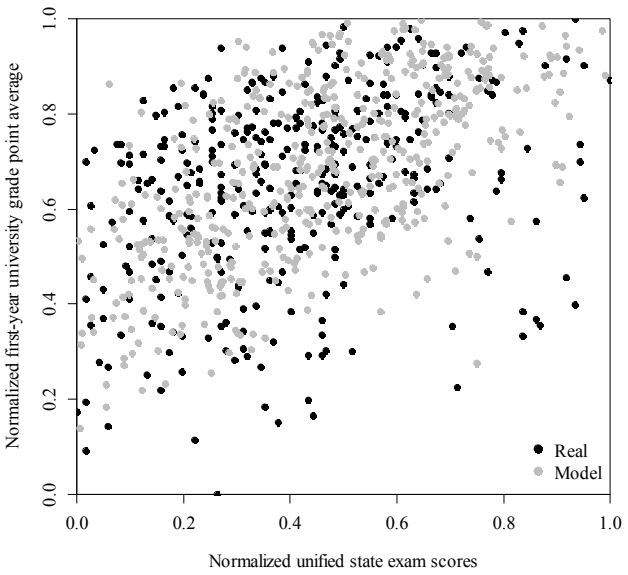

Fig. 1 Real and model scatter plot

As a source of real data the university database was taken. Sample of first-year students of the Faculty of Business was extracted. Rating system in this faculty enables to better differentiate the exam scores of weak students than strong compared to unified state exam scores.

VI. Simulation Results

A. Comparison of Weight Functions

First, the algorithm of orthogonal LOESS was studied. It is found that the objective function is not unimodal but it has as a rule one minimum. A behavior of loss function with different weight functions was compared. Fig. 2 illustrates loss functions graphs for example of a model sample. In both cases the number of neighbors is equal 200. The values of the objective function (17) are calculated at the point $\tilde{X}_0 = E(X) = \alpha/(\alpha+\beta)$. The findings indicate that loss function with rectangular weights is less stable than tricube weights function. This is explained by the fact that the rectangular weights do not change as smoothly. In terms of optimizing this is a disadvantage.

Second, predictive power of orthogonal LOESS was studied with different weight functions. It is evaluated based on root-mean-square differences $RMSE_{true}$ between predicted response values and the simulated values. The number of nearest neighbors (span) was varied from 200 to 450. Fig. 3 illustrates obtained results in the form of boxplots. The estimation results based on rectangular weight function considerably yield to results based on tricube weight function. They range widely, and medians of $RMSE_{true}$ show worse predictive power. Therefore by results of modeling the use of tricube weight function can be recommended. Further, this function is used for estimating orthogonal LOESS.

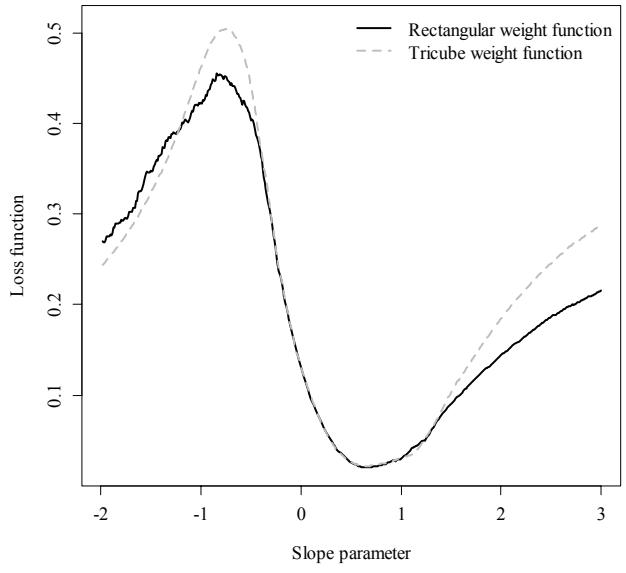

Fig. 2 Comparison of loss functions

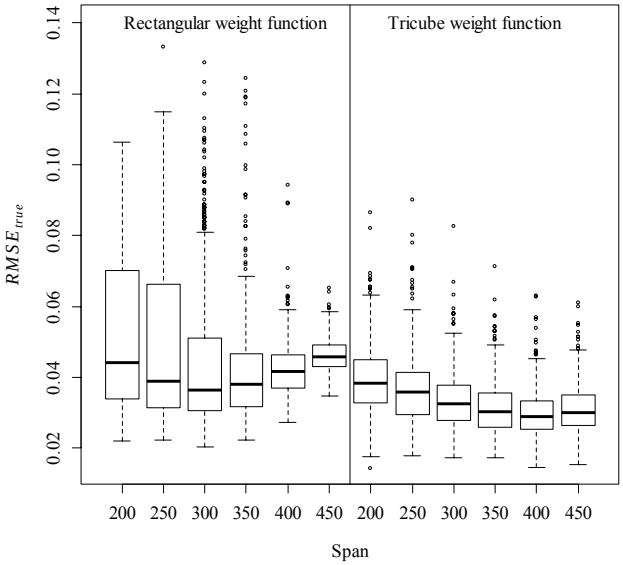

Fig. 3 Comparison of prediction errors

B. Comparison of Estimation Methods

The simulated data are used for testing of proposed algorithms. The number of knots was set equal to 51. The values of $\tilde{X}_j$ for orthogonal LOESS algorithm were set in a range from 0 to 1 on uniform grid by step 0.02. The number of nearest neighbors was varied from 40 to 450 by step 5. The smoothing parameter takes integer values from 1 to 65. Table I presents average optimal values of true root-mean-square error $RMSE^*_{true}$, the smoothing parameter λ^*, and the number of nearest neighbors k^* selected by the considered criteria. As an estimate of the average a median is taken. The values of interquartile range are given in parentheses (Table I).

TABLE I
Results of the Model Curve Estimation

Criterion	Orthogonal P-spline		Orthogonal LOESS	
	$RMSE^*_{true}$	λ^*	$RMSE^*_{true}$	k^*
$RMSE_{true}$	0.0429 (0.0087)	4 (4)	0.0354 (0.0061)	350 (191)
LOOCV	0.0584 (0.0140)	19 (58)	0.0553 (0.0158)	115 (80)
RMSE	0.0540 (0.0093)	24 (18)	0.0454 (0.0101)	450 (105)
MAE	0.0521 (0.0092)	17 (10)	0.0454 (0.0102)	447 (110)
MAPE	0.0553 (0.0098)	27 (24)	0.0456 (0.0101)	445 (115)

It is clear that use of the criterion of minimum of $RMSE_{true}$ provides true values of the smoothing parameters. In a really deviations from the true model are unknown. So the difference between the optimal values of the smoothing parameters obtained by criterion of minimum of $RMSE_{true}$ and other criteria indicates a quality of the fit.

Table I shows that orthogonal LOESS algorithm demonstrates the best results in terms of $RMSE_{true}$. Nevertheless using the smoothing parameter selection criteria cannot get its true value. The indexes of fit give the similar

results and have an oversmoothing effect. It is interesting that the use of cross-validation criterion leads to contrary results. In the case of penalized spline the effect oversmoothing occurs. By the usage of orthogonal LOESS algorithm inverse situation is observed.

VII. Conclusion

New algorithms for nonparametric estimation of errors-in-variables models are proposed. Influence of the weight function parameters on the prediction accuracy of orthogonal LOESS algorithm was investigated. Despite the fact that the rectangular weight function has a number of advantages (e.g. ease of calculation leave-one-out cross-validation) the usage of tricube weights leads to a smoother loss function. In addition, the use of tricube weights provides on the average smaller prediction error.

Furthermore the smoothing parameter selection criteria were investigated. Earlier [8] it was suggested to fit the smoothing parameter using cross-validation for initial step of the algorithm. This research has shown that it is really bad criterion in terms of the prediction accuracy. The use of usual fit indices gives better results. In general orthogonal LOESS algorithm is more perspective for confluence analysis.

Acknowledgment

The author would like to thank for financially supporting this research the Russian Foundation for Basic Research, project no. 14-07-31171 мол-а.

References

[1] J. L. Kobrin, B. F, Patterson, E. J. Shaw, K. D. Mattern, and S. M. Barbuti, "Validity of the SAT for predicting first-year college grade point average" (*College Board Research Report* No. 2008-5). New York: The College Board, 2008.

[2] J. J. Arneson, P. R. Sackett, and A. S. Beatty, "Ability-performance relationships in education and employment settings: critical tests of the more-is-better and the good-enough hypotheses," *Psychological Science*, vol. 22, no. 10, pp. 1336-1342, 2011.

[3] J. P. Marini, K. D. Mattern, and E. J. Shaw, "Examining the linearity of the PSAT/NMSQT®–FYGPA relationship" (*College Board Research Report* No. 2011-7). New York: The College Board, 2011.

[4] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*. New York: Cambridge university press, 2003.

[5] R. Blundell, X. Chen, and D. Kristensen, "Semi-nonparametric IV estimation of shape-invariant engel curves," *Econometrica*, vol. 75, no. 6, pp. 1613–1669, 2007.

[6] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, 1991.

[7] J. P. Buonaccorsi, *Measurement error: models, methods, and applications*. Boca Raton, London, New York: Chapman & Hall/CRC interdisciplinary statistics series, 2010.

[8] S. M. Berry, R. J. Carroll, and D. Ruppert "Bayesian smoothing and regression splines for measurement error problems," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 160-169, 2002.

[9] P. Pinson, H.A. Nielsen, H. Madsen, and T.S. Nielsen, "Local linear regression with adaptive orthogonal fitting for the wind power application," *Statistics and Computing*, vol. 18, no. 1, pp. 59-71, 2008.

[10] R. J. Carroll and D. Ruppert, "*The use and misuse of orthogonal regression in linear errors-in-variables models*," The American Statistician, vol. 50, no. 1, pp. 1-6, 1996.

[11] V. I. Denisov, A. Yu. Timofeeva, E. A. Khailenko, and O. I. Buzmakova "Robust estimation of nonlinear structural models," *Journal of Applied and Industrial Mathematics*, vol. 8, no. 1, pp. 28-39, 2014.

[12] W.S. Cleveland and S.J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596-610, 1988.

[13] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

[14] W. Härdle, *Applied nonparametric regression*. New York: Cambridge University Press, 1992.