The Korean Journal of Applied Statistics (2010)
23(4), 759–765

New Calibration Methods with Asymmetric Data

Sungsu Kim[1]

[1]Department of Statistics, Kyungpook National University

(Received May 2010; accepted June 2010)

Abstract

In this paper, two new inverse regression methods are introduced. One is a distance based method, and the other is a likelihood based method. While a model is fitted by minimizing the sum of squared prediction errors of y's and x's in the classical and inverse methods, respectively. In the new distance based method, we simultaneously minimize the sum of both squared prediction errors. In the likelihood based method, we propose an inverse regression with Arnold-Beaver Skew Normal(ABSN) error distribution. Using the cross validation method with an asymmetric real data set, two new and two existing methods are studied based on the relative prediction bias(RBP) criteria.

Keywords: Arnold-Beaver skew normal distribution, asymmetric data, inverse regression, calibration, relative prediction bias.

1. Introduction

Suppose one collects n paired observations on (x, y), where each y is observed for a fixed x. Let this stage be called the first phase. Later, suppose one observes one or more values of y's for a same value of x, but the value of the x is unknown. We call this stage the second phase. For example, n breaking distances of a car (y) are measured for pre-assigned driving speeds (x). This is the first phase. Later, suppose that a breaking distance of the car is observed but the corresponding driving speed is not recorded. This is the second phase. A calibration or inverse regression refers to making inferences about the corresponding value of driving speed (x) using one or more, say m, breaking distances (y's). In mathematical notation, the model is stated as

$$\begin{aligned} y_i &= a + bx_i + \epsilon_i, \quad i = 1, \ldots, n, \\ y_j &= a + b\widetilde{x} + \epsilon_j, \quad j = 1, \ldots, m, \end{aligned} \tag{1.1}$$

where a and b are slope and intercept parameters, respectively, ϵ has zero mean, n and m represent the number of pairs collected in the first phase and the number of y's observed in the second phase, respectively. $\widetilde{x}$ denotes the unknown value of x to be predicted, using m observed y's in the second phase.

For the calibration problem, the classical method is introduced in Krutchkoff (1967). In this method, a and b are estimated by minimizing the sum of squared prediction errors of y's. That is, the

This research was supported by the Kyungpook National University research fund, 2010.
[1]Assistant Professor, Department of Statistics, Kyungpook National University, 1370 Sangyuk-Dong, Buk-Gu, Daegu 702-701, Korea. E-mail: sungsu@knu.ac.kr

objective function to be minimized is shown below.

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2. \tag{1.2}$$

After solving the fitted equation for x, $\widetilde{x}$ is predicted as

$$\hat{\widetilde{x}} = -\frac{\hat{b}}{\hat{a}} + \frac{\bar{y}}{\hat{b}}, \tag{1.3}$$

where $\hat{a}$ and $\hat{b}$ are the least square estimators(LSE) of a and b, and $\bar{y}$ is the sample mean of m y's from the second phase.

As a competitor of the classical method, Klutchkoff (1967) propose the inverse method, where the model is fitted by minimizing the sum of squared prediction errors of x's. So, the objective function to be minimized is given by

$$\sum_{i=1}^{n}(x_i - c - dy_i)^2, \tag{1.4}$$

where c and d are re-parameterizations of $-a/b$ and $1/b$, respectively. Using the fitted equation, we predict $\widetilde{x}$ as

$$\hat{\widetilde{x}} = \hat{c} + \hat{d}\bar{y}, \tag{1.5}$$

where $\hat{c}$ and $\hat{d}$ are the LSEs of c and d, and $\bar{y}$ is the same as above.

These two methods are discussed in Krutchkoff (1967, 1969), Martinelle (1970), Halperin (1970), Minder and Whitney (1975), Brown (1979), and Chow and Shao (1990). Krutchkoff (1967) concludes that, in the range of x's, the inverse method performs better than the classical method based on the mean squared error criteria. Krutchkoff (1969) adds to his findings that there are situations where the classical method is better in extrapolation (Krutchkoff, 1969). Martinelle (1970) evaluates Krutchkoff's conclusion theoretically and claims that it is true only for small samples and using more than one observed y's can reduce the advantage of the inverse method. In Halperin (1970), it is discussed that the classical method is preferred based on Pitman's (1937) closeness. Minder and Whitney (1975) and Brown (1979) use a likelihood analysis and integrated mean square error, respectively to conclude that both methods are closely optimal. Chow and Shao (1990) studied the probability that the ratio of the two estimates differs from the unity, and found that the probability increases as the ratio of the standard deviation of errors to the regression slope decreases.

Two new methods, distance based and likelihood based, of calibration are introduced in this paper. In a distance based method, distances between raw data and fitted values are utilized to find the estimates as done in the classical and the inverse methods. On the other hand, we use the likelihood of data to fit a model in the likelihood-based method. Motivation for developing the distance-based method is from the concept that the data set can be used for predicting both y and x. It is shown that the method is equivalent to using the perpendicular distances from each data point to a fitted line. A likelihood-based method is introduced in order to model an asymmetrically distributed y's. This method is motivated by the fact that many distributions encountered in practice are usually asymmetric. In this method, we assume that errors in (1.1) follow the Arnold-Beaver Skew Normal(ABSN) distribution. Using the cross validation technique applied to a skewed real data set, two new and two existing methods are compared based on the relative prediction bias criteria(RPB) and the relative absolute prediction error(RAPE).

2. Methods

2.1. A new distance based inverse regression

Having the inverse regression problem, it is noted that the data set from the first phase is used for the dual purposes: prediction of y given x and prediction of x given y. This motivates an attempt to estimate a and b by simultaneously minimizing both sums of prediction errors used in the classical and the inverse methods. Consequently, the objective function to be minimized with respect to a and b is given by

$$\sum_{i=1}^{n}(y_i - a - bx_i)^2 + \sum_{i=1}^{n}\left(x_i - \frac{y_i - a}{b}\right)^2. \tag{2.1}$$

After taking derivative of (2.1) with respect to a and setting it to zero, one can easily verify that the estimator of a is given by $\bar{y} - \hat{b}\bar{x}$. In order to get $\hat{b}$, first substitute $\bar{y} - b\bar{x}$ for a in (2.1), then write it in a vector form by stacking n observations:

$$(\mathbf{y} - \bar{\mathbf{y}} - b(\mathbf{x} - \bar{\mathbf{x}}))'(\mathbf{y} - \bar{\mathbf{y}} - b(\mathbf{x} - \bar{\mathbf{x}})) + \left(\mathbf{x} - \bar{\mathbf{x}} - \frac{1}{b}(\mathbf{y} - \bar{\mathbf{y}})\right)'\left(\mathbf{x} - \bar{\mathbf{x}} - \frac{1}{b}(\mathbf{y} - \bar{\mathbf{y}})\right), \tag{2.2}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \bar{\mathbf{y}} = \bar{y}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad \bar{\mathbf{x}} = \bar{x}\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

After substituting

$$\bar{X} = \mathbf{x} - \bar{\mathbf{x}}, \qquad \bar{Y} = \mathbf{y} - \bar{\mathbf{y}},$$

it follows that

$$\left(1 + \frac{1}{b^2}\right)(\bar{Y} - b\bar{X})'(\bar{Y} - b\bar{X}). \tag{2.3}$$

Differentiating (2.3) with respect to b and setting it equal to zero yields the quartic equation shown below.

$$\left(\bar{X}'\bar{X}\right)\hat{b}^4 - \left(\bar{X}'\bar{Y}\right)\hat{b}^3 + \left(\bar{X}'\bar{Y}\right)\hat{b} - \left(\bar{Y}'\bar{Y}\right) = 0. \tag{2.4}$$

A quartic equation solver can be found in the web; for example, one is found at http://www.1728.com/quartic.htm. After getting $\hat{b}$ and then $\hat{a}$, $\widetilde{x}$ is predicted as

$$\hat{\widetilde{x}} = -\frac{\hat{a}}{\hat{b}} + \frac{\bar{y}}{\hat{b}}. \tag{2.5}$$

where $\hat{a}$ and $\hat{b}$ are the LSEs of a and b. It is well known that the least squared estimators, $\hat{a}$ and $\hat{b}$, are asymptotic normally distributed.

The new method can be viewed as minimizing the sum of squared vertical and horizontal distances

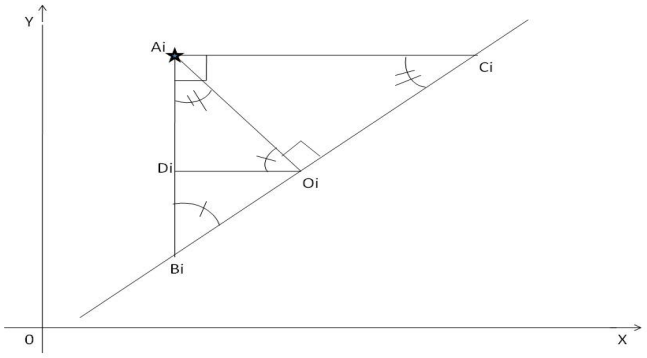

Figure 2.1. Graphical illustration of the new method.

from each data point to a fitted line. In the following, we show that minimizing the sum of squared distances is equivalent to minimizing the sum of the associated squared perpendicular distances. This is graphically illustrated in Figure 2.1. One can find five similar triangles in the figure. Examining two similar triangles denoted as $A_iB_iC_i$ and $D_iO_iA_i$, it is found that A_iB_i, A_iC_i and B_iC_i are similar sides to D_iO_i, D_iA_i and O_iA_i, respectively. Let A_i be a data point. Then, A_iB_i and A_iC_i represent vertical and horizontal distances from A_i to a fitted line, respectively, and A_iO_i is the associated perpendicular distance. In the following, it is shown that the sum of squared horizontal and vertical distances is a constant multiple of the sum of squared perpendicular distances.

$$\begin{aligned}\sum_{i=1}^{n}\left((A_iB_i)^2+(A_iC_i)^2\right) &= \sum_{i=1}^{n}(A_iB_i)^2\left(1+\frac{(A_iC_i)^2}{(A_iB_i)^2}\right)=\sum_{i=1}^{n}(A_iB_i)^2\left(1+\frac{(D_iA_i)^2}{(D_iO_i)^2}\right)\\ &=\sum_{i=1}^{n}\left(\frac{A_iB_i}{D_iO_i}\right)^2\left((D_iO_i)^2+(D_iA_i)^2\right)=c^2\sum_{i=1}^{n}\left(A_iO_i\right),\end{aligned} \tag{2.6}$$

where the last equal sign is due to $(A_iB_i)/(D_iO_i) = c$, for $i = 1, \ldots, n$, and c is some positive constant.

In the following sections, the method proposed in this section will be denoted as 'New' method.

2.2. Inverse regression with ABSN errors

Many distributions encountered in practice are usually asymmetric. This motivates us to develop a new calibration method with a skewed error distribution. The model is given by

$$y_i = a + bx_i + \epsilon_i, \tag{2.7}$$

where a and b are slope and intercept parameters, respectively, and ϵ_i's follow the ABSN distribution. The density of ϵ_i's is given by

$$f_{\epsilon_i}(\epsilon_i|\lambda_0,\lambda_1) = \frac{\exp\left(-\frac{\epsilon_i^2}{2}\right)\Phi\left(\lambda_0+\lambda_1\epsilon_i\right)}{\sqrt{2\pi}\Phi\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}, \quad i=1,\ldots,n, \tag{2.8}$$

where λ_0 and λ_1 are skewness parameters. ABSN distribution can model symmetric or asymmetric and unimodal or multimodal data (Arnold and Beaver, 2000). It is noted that it becomes the standard normal distribution when λ_0 and λ_1 are both equal to 0.

Having the above model, although location parameter is zero in (2.8), it is not centered at 0. Therefore, we find that the errors have non-zero conditional mean, which is equal to the bias given by

$$E\left(\epsilon_i|x_i\right) = \frac{\lambda_1}{\sqrt{1+\lambda_1^2}}\Lambda\left(\frac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right), \quad i = 1, \ldots, n, \tag{2.9}$$

where $\Lambda(\cdot)$ represents $\phi(\cdot)/\Phi(\cdot)$, the inverse Mill's ratio. Bias can be removed from the model by subtracting and adding the value of bias so that new errors can have a 0 conditional mean. Consequently, $\hat{a}$ is estimated with the added bias, and a correct value of $\hat{a}$ is given by the estimate of a minus the estimate of bias (Kim, 2009).

We use the maximum likelihood method to find the estimators since they possess a number of attractive asymptotic properties. The MLEs of a, b, λ_0 and λ_1 are solutions to the first order equations of the log likelihood function shown below with respect to a, b, λ_0 and λ_1.

$$\sum_{i=1}^{n}\left[\log\frac{\exp\left(-\dfrac{(y_i-a-bx_i)^2}{2}\right)\Phi\left(\lambda_0+\lambda_1(y_i-a-bx_i)\right)}{\sqrt{2\pi}\Phi\left(\dfrac{\lambda_0}{\sqrt{1+\lambda_1^2}}\right)}\right]. \tag{2.10}$$

After getting $\hat{a}$ and $\hat{b}$, $\hat{\widetilde{x}}$ is obtained as

$$\hat{\widetilde{x}} = -\frac{\hat{a}}{\hat{b}} + \frac{\overline{y}}{\hat{b}}, \tag{2.11}$$

where $\hat{a}$ and $\hat{b}$ are the MLEs of a and b. One may obtain these MLEs through a numerical computation. It is noted that computations of the works in this paper are easily facilitated using R. Since the ABSN density satisfies the regularity conditions, $\hat{a}$ and $\hat{b}$ are asymptotically normal with the asymptotic variance equal to the diagonal elements of the inverse of the information matrix. In the remaining sections, the method proposed in this section is labeled as 'ABSN' method.

2.3. Comparison of the methods

In this section, two existing and two new methods are compared based on the relative prediction bias(RPB) criteria. RPB for $\widetilde{x}$ is given by $E((\hat{\widetilde{x}}-\widetilde{x})/\widetilde{x})$. RPB measures the balanced amount by which each estimator over and underestimates so that smaller the RPB is more is that it will be correct on the average. As one of the referees suggested, one may also consider $E|(\hat{\widetilde{x}}-\widetilde{x})/\widetilde{x}|$ in comparing the methods, which may be called the relative absolute prediction error(RAPE). When using the cross validation method, the overall estimated relative prediction bias(OERPB) and the overall relative absolute prediction bias(OERAPE) are given by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{x}_{(i)}-x_i}{x_i} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{x}_{(i)}-x_i}{x_i}\right|,$$

respectively, where $\hat{x}_{(i)}$ is obtained using the data set with x_i omitted.

Data set, which is found in the R data base, consists of 66 plasma concentrations of Indomethicin

Table 2.1. OERPB using the 66 plasma concentrations data set.

	Methods			
	Classical	Inverse	New	ABSN
OERPB	−1.82	1.81	−1.63	−1.00

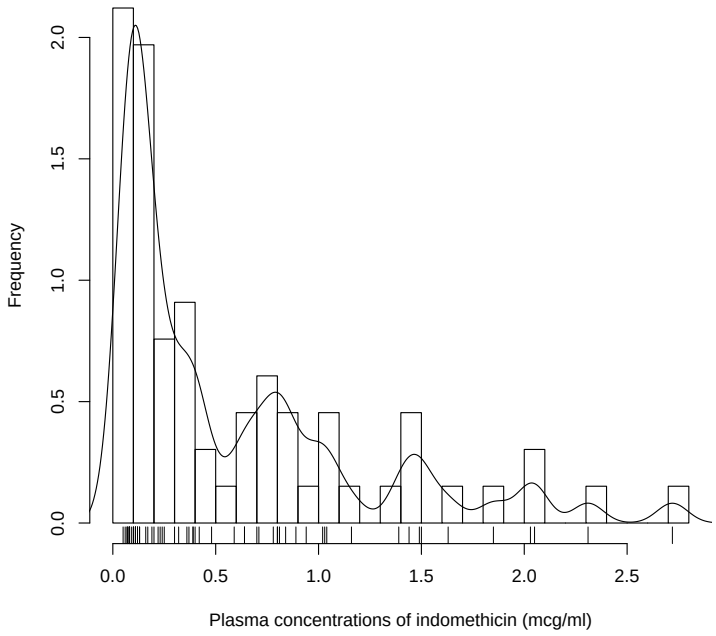

Figure 2.2. Density plot of 66 plasma concentrations of Indomethicin.

(mcg/ml) for different times (hr) at which blood samples were drawn. A smoothed density plot of the 66 plasma concentrations is displayed in Figure 2.2. It shows that the data set is clearly skewed to the right. In Table 2.1, the OERPBs of the existing methods and new methods are presented.

3. Discussion

In this paper, two new methods in calibration were proposed. It was shown that the methods are easy to be estimated and implemented using R. For the asymmetric data set containing 66 blood samples and plasma concentrations, it is found that two new methods performed well for the data set, of which the ABSN method yielded the smallest OERPB as −1.00. Therefore, it is suggested that, for an asymmetrically distributed data set, two new methods are preferred to two existing methods based on the RPB criterion.

References

Arnold, B. and Beaver, R. (2000). Hidden truncation models, *Sankhya*, **62**, 23–35.

Brown, G. H. (1979). An optimization criterion for linear inverse estimation, *Technometrics*, **21**, 727–736.

Chow, S. and Shao, J. (1990). On the difference between the classical and inverse methods of calibration, *Applied Statistics*, **39**, 219–228.

Halperin, M. (1970). On inverse estimation in linear regression, *Technometrics*, **12**, 595–601.

Kim, S. (2009). *Inverse Circular Regression with Possibly Asymmetric Error Distribution*, PhD Dissertation, University of California, Riverside.

Krutchkoff, R. G. (1967). Classical and inverse regression methods of calibration, *Technometrics*, **9**, 425–439.

Krutchkoff, R. G. (1969). Classical and inverse regression methods of calibration in extrapolation (in notes), *Technometrics*, **11**, 605–608.

Martinelle, S. (1970). On the choice of regression in linear calibration, *Technometrics*, **12**, 157–161.

Minder, C. E. and Whitney, J. B. (1975). A likelihood analysis of the linear calibration problem, *Technometrics*, **17**, 463–471.

Pitman, E. (1937). The closest estimates of statistical parameters, *Proceedings of the Cambridge Philosophical Society*, **33**, 212–222.