Pergamon

0731-7085(94)00057-3

Journal of Pharmaceutical & Biomedical Analysis, Vol. 12, No. 10, pp. 1265–1269, 1994
Copyright © 1994 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0731-7085/94 $7.00 + 0.00

Fitting a straight line when both variables are subject to error: pharmaceutical applications

TAPON ROY

Boehringer Ingelheim Pharmaceuticals, Inc. 900 Ridgebury Road, Ridgefield, CT 06877, USA

Abstract: In many pharmaceutical applications one postulates a linear relationship between variables. The usual linear least-squares methods are appropriate when the values of the independent variable are constants, and the dependent variable is subject to error. When both variables are subject to error, as in assay validation, calibration, and general correlation, the measurement error model (also called errors-in-variables) should be used especially when independent variable error is appreciable. In this paper, the theoretical properties of errors-in-variables methods are demonstrated with examples, and a technique for assessing the variability of parameter estimates without normality assumptions is presented. Robust methods resistant to outliers and not requiring normality assumptions, are also described.

Keywords: *Regression; assay validation; calibration; correlation; measurement error; robust methods.*

Introduction

When analysing pharmaceutical data, one often postulates a linear relationship between variables. In quantifying a dose–response relationship or evaluating a stability profile, the usual linear regression methods can be used, since the independent variable (X) consists of constant values at which corresponding measurements of the dependent variable (Y) are made. In ordinary linear regression, only Y is assumed subject to error. In this case it is well-known that ordinary least-squares (under the customary additional assumptions) gives the best linear unbiased estimates of the slope and intercept parameters.

In many other situations, such as assay validation, calibration, and general correlation, both X and Y are subject to error and ordinary least squares may be inappropriate since the absolute magnitude of the slope is underestimated. Both X and Y can be subject to error from two sources. The first source is random error associated with the assumption that Y and possibly X are random. The second source is the measurement error made in determining the values of X and Y. When both X and Y are assumed random the model is called *structural*. We will consider the most general case — the structural model where both X and Y may be subject to measurement error. When both variables are subject to error, the slope estimated by ordinary least squares is biased toward zero [1]. Thus, the measurement error model (also called errors-in-variables) should be used to estimate the slope and intercept [1]. The variability of the parameter estimates, and of the correlation coefficient, is also assessed, both with and without assuming normality for X and Y. Robust methods, resistant to outliers as well as not needing normality assumptions [2–9] are also useful in many instances.

Methods

Following the derivation and notation in Fuller [1], consider parameter estimation for the measurement error model, given knowledge of the relative magnitude of the error variances of the two variables. The model is

$$\begin{gathered} y_i = \beta_0 + \beta_1 x_i, \\ (Y_i, X_i) = (y_i, x_i) + (e_i, u_i), \end{gathered} \tag{1}$$

where (Y_i, X_i) is observed, y_i is the true value of the dependent variable, x_i is the true value of the independent variable, and (e_i, u_i) are the measurement errors, assumed to be normally distributed. The terms 'independent variable' and 'dependent variable' are sometimes still used although the model is symmetric in x and y for $\beta_1 \neq 0$.

When the ratio

$$\delta = \sigma_{uu}^{-1}\sigma_{ee} \tag{2}$$

is known, the model is the classical errors-in-variables model.

The estimator of the slope is

$$\hat{\beta}_1 = \frac{m_{YY} - \delta m_{XX} + [(m_{YY} - \delta m_{XX})^2 + 4\delta m_{XY}^2]^{1/2}}{2m_{XY}} \tag{3}$$

where m_{YY}, m_{XY}, and m_{XX} are the sample variance of y, covariance, and variance of x, respectively, and δ is the estimated ratio of y to x error variances, given a default value of one.

The estimator of the intercept is then

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \tag{4}$$

where $(\bar{Y},\bar{X})$ are the sample means.

The estimated standard error of $\hat{\beta}_1$ is

$$\text{SE}\{\hat{\beta}_1\} = \{(n-1)^{-1}\hat{\sigma}_{xx}^{-2}[\hat{\sigma}_{xx}s_{vv} + \hat{\sigma}_{uu}s_{vv} - \hat{\sigma}_{uv}^2]\}^{1/2} \tag{5}$$

where

$$s_{vv} = (n-2)^{-1})\,(n-1)\,(\delta + \hat{\beta}_1^2)\hat{\sigma}_{uu} \tag{6}$$

$$\hat{\sigma}_{xx} = (2\delta)^{-1}\{[(m_{YY} - \delta m_{XX})^2 + 4\delta m_{XY}^2]^{1/2} - (m_{YY} - \delta m_{XX})\} \tag{7}$$

$$\hat{\sigma}_{uu} = (2\delta)^{-1}\{m_{YY} + \delta m_{XX} - [(m_{YY} - \delta m_{XX})^2 + 4\delta m_{XY}^2]^{1/2}\} \tag{8}$$

$$\hat{\sigma}_{uv} = -\hat{\beta}_1\hat{\sigma}_{uu} \tag{9}$$

and

$$\text{SE}\{\hat{\beta}_0\} = [n^{-1}s_{vv} + \bar{X}^2\,\hat{V}\{\hat{\beta}_1\}]^{1/2} \tag{10}$$

where

$$\hat{V}\{\hat{\beta}_1\} = (n-1)^{-1}\hat{\sigma}_{xx}^{-2}[\hat{\sigma}_{xx}s_{vv} + \hat{\sigma}_{uu}s_{vv} - \hat{\sigma}_{uv}^2] \tag{11}$$

and n is the sample size.

When X and Y are normally distributed, for $n \geqslant 3$, a $1 - \alpha$ confidence interval for the correlation coefficient, ρ, is given by

$$\tanh[\operatorname{arctanh}(\rho) - z_{\alpha/2}(n-3)^{1/2}] \leqslant \rho \leqslant \tanh[\operatorname{arctanh}(\rho) + z_{\alpha/2}(n-3)^{1/2}] \tag{12}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ probability point of the standard normal distribution. This confidence interval is based on Fisher's Z transformation [10].

If the variables are not normally distributed, then the standard errors for the parameter estimates and the confidence interval for the correlation coefficient presented above, may not be valid. An alternative method, the bootstrap [11], can be used to estimate standard errors and confidence intervals when nonnormality is present. We propose the following procedure: since both X and Y were subject to error, (Y_i, X_i) pairs were resampled, with replacement, from the original data and the errors-in-variables parameter estimates and the correlation coefficient computed for each resample. Bootstrap standard errors and confidence intervals were then calculated from the empirical distributions. Since the resampling was done with replacement there was a very slight chance of obtaining anomalously extreme estimates in the tails of the bootstrap distributions. Thus, the distribution of the correlation coefficient was trimmed 1% at each end before the confidence interval was calculated. Instead of standard errors, the median absolute deviations of the parameter estimates are presented. These are based on the untrimmed distributions and are defined as the median of the absolute values of the differences between each element of the distribution and the median of the parameter distribution. More complex methods for adjusting bootstrap confidence intervals for the correlation coefficient are also available [12].

When outliers are present in the data, the least-squares based regression techniques (ordinary and errors-in-variables) can be seriously affected. If nonnormality is also present, interval estimation can be especially unreliable since the parameter estimates are being shifted due to the outliers in addition to the violation of the normality assumption by, possibly, other causes. When only Y is subject to error, a method such as least absolute value, or L_1, regression is quite resistant to outliers. When both X and Y are subject to errors, techniques based on using the median of all possible defined and finite pairwise slopes [2, 3, 5] or the least median of squares [6, 8, 9] have been proposed.

Results and Discussion

To demonstrate the methods, several of the hydrophobicity measures considered in [13] were analysed to determine linear relationships between variables. The data appear in Table 1 of [13]. In Table 1 of this paper, ordinary and errors-in-variables regression estimates are presented, using the logarithm of the experimental *n*–octanol–water partition coefficient (log *P*) as the dependent variable, and the logarithm of the LC capacity factor (k') determined on immobilized artificial membrane with acetonitrile–pH 7.00 buffer (20:80, v/v) as the independent variable. In Table 2 log *P* is fit to a k' value similar to the one used in Table 1, but with acetonitrile–pH 7.00 buffer (25:75, v/v). In Table 3, the k' value used in Table 2 is fit to the logarithm of the k' from a deactivated hydrocarbonaceous silica column, normalized to 0% organic modifier in mobile phase. In all tables, all solutes were used, and the nomenclature is dependent (*Y*) vs independent (*X*). The estimated ratio of error variances, δ, is set to one.

As predicted in theory, the slopes estimated in Tables 1–3 from errors-in-variables exceed, in absolute magnitude, those obtained from ordinary least squares. The effect is especially pronounced in Tables 1 and 2. Since no departure from normality was evident for the variables in Tables 1 and 2, the usual normal theory standard errors and confidence interval were used. In Table 3, some moderate departure from normality for the variables was detected. (Note that though log k'_{75} was used in both Tables 2 and 3, many more values could be used in Table 3). Thus, the errors-in-variables parameter standard errors and confidence interval were resampling-based. For comparison, the normal theory standard error for the errors-in-variables intercept was 0.134, and the standard error for the slope was 0.037. The correlation coefficient was not redefined for the errors-in-variables case, and from Table 3 it is clear that the bootstrap confidence interval is considerably narrower than the interval based on the normal theory based Fisher *Z*-transformation.

An illustrative example of the value of robust methods is the comparison of two analytical methods TOA and BGE for measuring the packed cell volume (PCV), or hematocrit [9]. The data are in the Appendix of [9]. To test proportional and additive accuracy, one can

Table 1
Comparison of ordinary linear regression with errors-in-variables linear regression, log *P* vs log k'_{80} (all solutes). Normality assumption is made for both variables

	n	Intercept, $\hat{\beta}_0$ (SE)	Slope, $\hat{\beta}_1$ (SE)	Correlation coefficient (95% confidence interval)
Ordinary linear regression	19	2.299 (0.373)	1.527 (0.264)	0.814 (0.572, 0.926)*
Errors-in-variables linear regression	19	1.510 (0.512)	2.119 (0.369)	

* Confidence interval based on Fisher's *Z*-transformation.

Table 2
Comparison of ordinary linear regression with errors-in-variables linear regression, log *P* vs log k'_{75} (all solutes). Normality assumption made for both variables

	n	Intercept, $\hat{\beta}_0$ (SE)	Slope, $\hat{\beta}_1$ (SE)	Correlation coefficient (95% confidence interval)
Ordinary linear regression	19	2.405 (0.391)	1.699 (0.324)	0.786 (0.517, 0.914)*
Errors-in-variables linear regression	19	1.438 (0.580)	2.551 (0.491)	

* Confidence interval based on Fisher's *Z*-transformation.

Table 3
Comparison of ordinary linear regression with errors-in-variables linear regression, log k'_{75} vs log k'_{w} (all solutes). No normality assumption made for errors-in-variables variability measures

	n	Intercept, $\hat{\beta}_0$ (SE)	Slope, $\hat{\beta}_1$ (SE)	Correlation coefficient (95% confidence interval)
Ordinary linear regression	29	−0.229 (0.130)	0.389 (0.036)	0.902 (0.799, 0.953)*
Errors-in-variables linear regression	29	−0.271 (0.108)‡	0.401 (0.029)‡	0.902 (0.826, 0.954)†

* Confidence interval based on Fisher's *Z*-transformation.
† Confidence interval is based on the trimmed bootstrap, B = 2000.
‡ Median absolute deviations from the untrimmed bootstrap, B = 2000.

Table 4
Comparison of hematocrit (PCV) methods, BGE vs TOA

	n	Intercept β_0 (SE)	Slope, β_1 (SE)
Ordinary linear regression	112	0.163 (0.035) (0.093, 0.232)*	0.586 (0.097) (0.395, 0.778)*
Errors-in-variables linear regression	112	−0.116 (0.081) (−0.276, 0.045)†	1.365 (0.226) (0.918, 1.812)†
Robust linear regression (Theil-Sen)	112	0.013 (−0.014, 0.045)‡	1.0 (0.909, 1.081)‡

*95% Confidence interval, normal theory based.
†Asymptotically normal theory based 95% confidence interval.
‡Order-statistic based 95% confidence interval, large sample approximation.

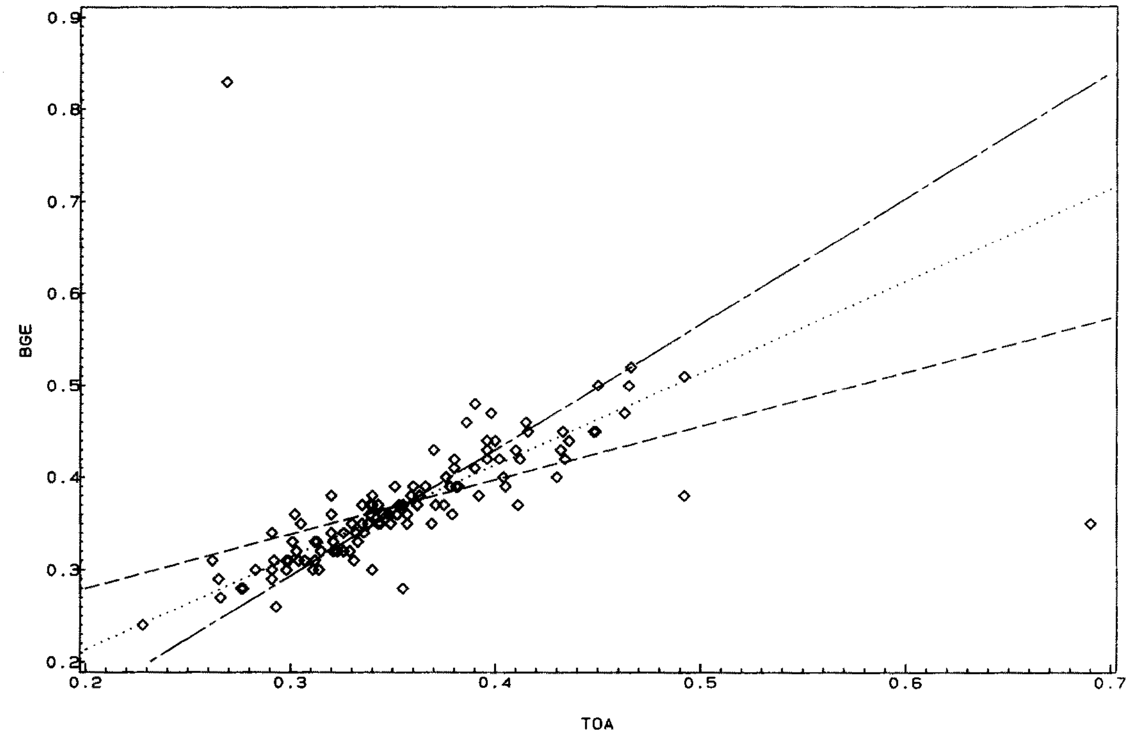

Figure 1
Comparison of hematocrit (PCV) techniques: BGE (Y) vs TOA (X). (----) Ordinary least squares; (— — — —) errors-in-variables; (· · · · ·) robust regression.

test whether or not the slope and intercept are significantly different from one and zero, respectively. The data include several serious influential outliers, and in Table 4 and Fig. 1 the effect on the parameter estimates and fitted lines is substantial. The ordinary least squares estimates are most severely compromised. The errors-in-variables slope estimate indicates the marked underestimation of the slope by ordinary least-squares, though the errors-in-variables method is also affected by outliers. The robust linear regression [2–4] results show the effect of the outliers on the ordinary and errors-in-variables methods.

Note that the robust method used is valid when both variables are subject to error [3]. The errors-in-variables and robust techniques both lead to the conclusion that there are no significant proportional or additive biases between the hematocrit methods since the confidence intervals for the slope and intercept include 1 and 0, respectively. The ordinary linear regression results, however, indicate differences in both proportional and additive accuracy. Thus, use of ordinary least squares when both variables are subject to error, and/or when outliers are present can lead to erroneous conclusions. Additionally, the robust method is resistant to departures from normality by any cause.

There are other robust methods [5, 6] that are even more resistant to outliers than the one used in this paper, but they can exhibit local instability [8] or may be computationally burdensome [7]. Also, standard errors or confidence intervals have not been developed for these methods.

These examples reinforce the theoretical results comparing regressions techniques when both variables are subject to error, outliers are present, and provide methods to assess variability when normality assumptions might be untenable.

There are alternative approaches proposed for errors-in-variables estimation [1, 14–16], but the method described here is realistic and flexible. In particular, the value of δ can be adjusted to account for the relative importance of the error variances. A value of one gives equal importance to horizontal and vertical deviations. For example, a value of 4 would give four times the importance to the vertical deviations, and a value of 1/5 gives five-fold importance to the horizontal distances. Given knowledge of how precisely the variables were measured, one could finetune δ for each specific case. Measurement error models can also be used when there are multiple independent variables and for nonlinear regression models.

Computations were performed using the SAS system [17]. Bootstrap methods were applied using a modification of a resampling program [18] using data pairs rather than residuals since both variables were considered random.

Acknowledgements — The author wishes to thank the reviewers for their helpful comments and the BIPI Document Publishing Group for their expert word processing of the manuscript.

References

[1] W.A. Fuller, *Measurement Error Models*. John Wiley, New York (1987).

[2] H. Theil, *Nederl. Akad. Wetensch. Proc.* **53**, 386–392, 521–525, 1397–1412 (1950).

[3] P.K. Sen, *J. Am. Statist. Assoc.* **63**, 1379–1389 (1968).

[4] M. Hollander and D.A. Wolfe, *Nonparametric Statistical Methods*. John Wiley, New York (1973).

[5] A.F. Siegel, *Biometrika* **69**, 242–244 (1982).

[6] P.J. Rousseeuw, *J. Am. Statist. Assoc.* **79**, 871–880 (1984).

[7] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*. John Wiley, New York (1986).

[8] T.P. Hettmansperger and S.J. Sheather, *Am. Statist.* **46**, 79–83 (1992).

[9] U. Feldmann, *Eur. J. Clin. Chem. Clin. Biochem.* **30**, 405–414 (1992).

[10] R.A. Fisher, *Metron* **1**, 3–32 (1921).

[11] B. Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia (1982).

[12] P. Hall, M.A. Martin and W.R. Schucany, *J. Statist. Comput. Simul.* **33**, 161–172 (1989).

[13] R. Kaliszan, A. Kaliszan and I.W. Wainer, *J. Pharm. Biomed. Anal.* **11**, 505–511 (1993).

[14] C. Spiegelman, *Ann. Statist.* **7**, 201–206 (1979).

[15] R.H. Ketellapper and A.E. Ronner, *Metrika* **31**, 33–41 (1984).

[16] A.S. Whittemore, *Am. Statist.* **43**, 226–228 (1989).

[17] SAS Institute, Inc., *SAS/STAT User's Guide, Version 6, Fourth Edition, Volumes 1 and 2*. SAS Institute, Inc., Cary, NC (1989).

[18] R.T. Carson, in *Proceedings of the Tenth Annual SAS Users Group International Conference*, pp. 1064–1069. SAS Institute, Inc., Cary, NC (1985).

[Received for review 11 February 1994; revised manuscript received 5 March 1994]