Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Analytical and monte carlo comparisons of six different linear least squares fits

Gutti Jogesh Babu & Eric D. Feigelson

To cite this article: Gutti Jogesh Babu & Eric D. Feigelson (1992) Analytical and monte carlo comparisons of six different linear least squares fits, Communications in Statistics - Simulation and Computation, 21:2, 533-549

To link to this article: https://doi.org/10.1080/03610919208813034

COMMUN. STATIST. -SIMULA., 21(2), 533-549 (1992)

ANALYTICAL AND MONTE CARLO COMPARISONS OF SIX DIFFERENT LINEAR LEAST SQUARES FITS

Gutti Jogesh Babu

Department of Statistics
219 Pond Laboratory
Pennsylvania State University
University Park PA 16802

Eric D. Feigelson

Dept. of Astronomy & Astrophysics
525 Davey Laboratory
Pennsylvania State University
University Park PA 16802

Keywords and Phrases: Tully-Fisher relation; orthogonal regression; reduced major axis; linear regression; variance estimation; cosmic distance scale.

ABSTRACT

For many applications, particularly in allometry and astronomy, only a set of correlated data points (x_i, y_i) is available to fit a line. The underlying joint distribution is unknown, and it is not clear which variable is 'dependent' and which is 'independent'. In such cases, the goal is an intrinsic functional relationship between the variables rather than E(Y|X), and the choice of least-squares line is ambiguous. Astronomers and biometricians have used as many as six different linear regression methods for this situation: the two ordinary least-squares (OLS) lines, Pearson's orthogonal regression, the OLS-bisector, the reduced major axis and the OLS-mean. The latter four methods treat the X and Y variables symmetrically. Series of simulations are described which compared the accuracy of regression estimators and their asymptotic variances for all six procedures. General relations between the regression slopes are also

obtained. Among the symmetrical methods, the angular bisector of the OLS lines demonstrates the best performance. This line is used by astronomers and might be adopted for similar problems in biometry.

INTRODUCTION

Long-standing controversy exists within various research communities over which of several alternative procedures represents the 'best' linear fit to bivariate data of the form (x_i,y_i). The difficulty arises when nothing is known about the underlying joint distribution, because the variables under study are complex and poorly understood. Examples of such problems include mammalian metabolism and surface area used in Kleiber's law in allometry, and galaxy rotation and luminosity used in the Tully-Fisher relation in astronomy. In these situations and others in social sciences such as economics, the classification of variables as *dependent* and *independent* is unclear or arbitrary. The goal of such studies is often a quantitative descrption of the underlying functional relationship between the variables, rather than the conditional expectation of one variable given the other. Clear statistical criteria such as maximum likelihood can not be applied. The search for methods which mathematically treat both variables symmetrically, thereby avoiding the distinction between dependent and independent variables, has led to the consideration of several alternative least-squares procedures. This paper and a companion study (Isobe *et al.* 1990; henceforth IFAB) provide a thorough account of these methods, with self-consistent analytical and numerical error analysis.

The problem was first clearly stated by Pearson (1901) who, noting that ordinary least squares (OLS) gives different lines when the X and Y variables are switched, proposed a regression that is insensitive to such changes. He states "the term 'best fit' is really arbitrary ... a good fit will clearly be obtained if we make the sum of the squares of the perpendiculars from the system of points upon the line ... a minimum". This is now known as Pearson's major axis or orthogonal regression. Linnik (1961) showed that

it is the unique line that minimizes the moment of inertia of the system. Scientists have been uncertain when it should be used. In chemometrics, for example, some monographs (Shorter 1982) recommend the use of the standard OLS(Y|X), where X is the independent variable. Others consider this choice 'arbitrary' and state that orthogonal regression is more logical under certain conditions (Massart *et al.* 1978).

Biometricians Kermack and Haldane (1950) raised a related issue for situations where the underlying probability structure is unknown and "biological variability" is the dominant source of scatter. In these cases, "the conventionally used regression lines are quite unsuitable, since here the terms 'dependent variable' and 'independent variable' have no real meaning". They, and prior to their work astronomer Strömberg (1940), proposed another regression line treating X and Y symmetrically, with slope equal to the geometric mean of the two OLS slopes. This line, known as the 'reduced major axis' or 'geometric mean' regression line, is scale-invariant. Allometers have extensively discussed the relative merits of orthogonal regression, the reduced major axis and the two OLS lines (*e.g.* Sokal and Rohlf 1981; Jungers 1985).

More recently, astronomers have proposed two additional regression procedures that treat the variables symmetrically: the line that bisects the angle formed by the two OLS lines (henceforth called the OLS-bisector; *e.g.* Rubin *et al.* 1980; Pierce and Tully 1988); and the line whose slope is the arithmetic mean of the two OLS slopes (henceforth called the OLS-mean; *e.g.* Aaronson *et al.* 1986). Together with the orthogonal regression and reduced major axis, modern astronomers thus use as many as four least-squares procedures treating the variables symmetrically for bivariate data sets (see IFAB).

Among astronomers, there is a common misconception that the four symmetric lines can be used interchangeably to represent the same population relation. This confusion is most apparent in studies of the 'cosmic distance scale', large-scale motions in the universe, and related topics (*e.g.* Rowan-Robinson 1985; Rubin and Coyne 1989). Differences in regression methods on

similar data may be responsible for a portion of the long-standing controversy over the value of Hubble's constant, which quantifies the recession rate of the galaxies. Various astronomers using the spiral galaxy Tully-Fisher relation, an important step in the cosmic distance scale, have used the standard OLS(Y|X) (Freedman 1990), OLS(X|Y) (Tully 1988), OLS-bisector (Pierce and Tully 1988), and OLS-mean (Aaronson *et al.* 1986). The resulting values of Hubble's constant, and related astronomical phenomena like galaxy streaming, derived by these researchers may be incompatible due to their use of different regression methods. Also, the error analyses in such studies, usually incorrectly based on standard OLS(Y|X) variances (*e.g.* Bevington 1969), will be wrong.

A similar uncertainty also exists in the biometrical community regarding the merits of the lines treating the variables symmetrically (see the discussions in Sokal and Rohlf 1981, and papers within Jungers 1985). The debate partly concerns the mathematical properties of the procedures, and partly the scientific purpose of the experiment being analyzed. For example, some researchers argue that the method chosen should depend on the goal of the experiment. Symmetric methods might be used when comparing data to theory, and the standard OLS (Y|X) adopted when predicting new values of Y from measured values of X. Specific examples requiring these methods are given in IFAB.

The present effort attempts to clarify some of the confusion. The expressions for the regression coefficients and their variances (Table 1 below) are calculated using the delta method (Cramér 1946; Billingsley 1986) in IFAB. In this paper, we first establish mathematically the relationships between the various least squares linear slopes. Monte Carlo simulations that test the large- and small-sample performances of our coefficient and variance estimators are presented. In the final analysis, the angular bisector is perhaps the most desirable of the symmetric methods.

RELATIONSHIPS BETWEEN THE SLOPES

We address the following six regression lines:

OLS(Y|X), the standard ordinary least squares line which minimizes the sum of squares of vertical residuals.

OLS(X|Y), the 'inverse' regression line which minimizes the sum of squares of horizontal residuals (see Krutchkoff 1967).

OLS-bisector, which bisects the angle formed by the two OLS lines.

OR (orthogonal regression), which minimizes the sum of squares of perpendicular distances.

RMA (reduced major axis), whose slope is the geometric mean of the two OLS slopes.

OLS-mean, whose slope is the arithmetic mean of the two OLS slope.

If the underlying population is characterized by standard deviations σ_x and σ_y in X and Y, and correlation coefficient $\rho \neq 0$, then the population slopes of these six lines can be expressed in simple analytical form (see IFAB):

$$\text{OLS(Y | X)} \quad \beta_1 = \rho\sigma_y/\sigma_x \tag{1}$$

$$\text{OLS(X | Y)} \quad \beta_2 = \sigma_y/\rho\sigma_x \tag{2}$$

$$\text{OLS} - \text{bisector} \quad \beta_3 = \frac{\rho}{1+\rho^2}\{\frac{\sigma_y^2-\sigma_x^2}{\sigma_x\sigma_y} + [(\frac{\sigma_x}{\sigma_y})^2 + \rho^2 + \rho^{-2} + (\frac{\sigma_y}{\sigma_x})^2]^{1/2}\} \tag{3}$$

$$\text{OR} \quad \beta_4 = \frac{1}{2\rho\sigma_x\sigma_y}\{\sigma_y^2 - \sigma_x^2 + [(\sigma_y^2-\sigma_x^2)^2 + 4\rho^2\sigma_x^2\sigma_y^2]^{1/2}\} \tag{4}$$

$$\text{RMA} \quad \beta_5 = \frac{\sigma_y}{\sigma_x} sign(\rho) \tag{5}$$

$$\text{OLS} - \text{mean} \quad \beta_6 = \tfrac{1}{2}(\beta_1+\beta_2) = \tfrac{1}{2}(\rho+\rho^{-1})\sigma_y/\sigma_x. \tag{6}$$

These formulas show clearly that the six lines have different dependencies on the underlying population correlation ρ, and thus should not be considered estimators of the same quantity. We also define the following quantities: $\overline{x} = 1/n \sum_{i=1}^n (x_i)$, $\overline{y} = 1/n \sum_{i=1}^n (y_i)$, $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$, $S_{yy} = \sum_{i=1}^n (y_i - \overline{y})^2$, and $S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$.

TABLE I

Formulae for Six Linear Regression Slopes

Method	Expression for Slope	The Estimate of the Variance of the Slope $\widehat{\mathrm{Var}}(\hat{\beta}_i)$
OLS(Y|X)	$\hat{\beta}_1 = S_{xy}/S_{xx}$	$[\sum_{i=1}^{n}(x_i-\overline{x})^2(y_i-\hat{\beta}_1 x_i-\overline{y}+\hat{\beta}_1\overline{x})^2]/S_{xx}^2$
OLS(X|Y)	$\hat{\beta}_2=S_{yy}/S_{xy}$	$[\sum_{i=1}^{n}(y_i-\overline{y})^2(y_i-\hat{\beta}_2 x_i-\overline{y}+\hat{\beta}_2\overline{x})^2]/S_{xy}^2$
OLS-bisector	$\hat{\beta}_3 = (\hat{\beta}_1+\hat{\beta}_2)^{-1}[\hat{\beta}_1\hat{\beta}_2 - 1 + [(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)]^{\frac{1}{2}}]$	$[\hat{\beta}_3^2/(\hat{\beta}_1+\hat{\beta}_2)^2(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)]\,[(1+\hat{\beta}_2^2)^2\widehat{\mathrm{Var}}(\hat{\beta}_1)) + 2(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) + (1+\hat{\beta}_1^2)^2\widehat{\mathrm{Var}}(\hat{\beta}_2)]$
Orthogonal regression	$\hat{\beta}_4 = \frac{1}{2}[(\hat{\beta}_2-\hat{\beta}_1^{-1}) + sign(S_{xy})[4+(\hat{\beta}_2-\hat{\beta}_1^{-1})^2]^{\frac{1}{2}}]$	$\hat{\beta}_4^2[\hat{\beta}_1^{-2}\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) + \hat{\beta}_1^2\widehat{\mathrm{Var}}(\hat{\beta}_2)]/ [4\hat{\beta}_1^2+(\hat{\beta}_1\hat{\beta}_2-1)^2]$
Reduced major axis	$\hat{\beta}_5 = sign(S_{xy})(\hat{\beta}_1\hat{\beta}_2)^{\frac{1}{2}}$	$\frac{1}{4}[(\hat{\beta}_2/\hat{\beta}_1)\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) + (\hat{\beta}_1/\hat{\beta}_2)\widehat{\mathrm{Var}}(\hat{\beta}_2)]$
OLS-mean	$\hat{\beta}_6 = \frac{1}{2}(\hat{\beta}_1+\hat{\beta}_2)$	$\frac{1}{4}[\widehat{\mathrm{Var}}(\hat{\beta}_1) + \widehat{\mathrm{Var}}(\hat{\beta}_2) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2)]$

An estimate of covariance term is given by

$$\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) = \{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})\,[y_i-\overline{y}-\hat{\beta}_1(x_i-\overline{x})]\,[y_i-\overline{y}-\hat{\beta}_2(x_i-\overline{x})]\}\ /(\hat{\beta}_1 S_{xx}^2)$$

Now, let $\hat{\beta}_i$ denote the corresponding estimates of β_i based on a given data set $(x_i, y_i), \ldots, (x_n, y_n)$. Table 1 presents the six sample slope coefficients $\hat{\beta}_i$ and their variances $\widehat{Var}(\hat{\beta}_i)$. The variances for $\hat{\beta}_1, \ldots, \hat{\beta}_5$ are derived in Appendix A of IFAB using the 'delta method', and thus rely on the central limit theorem. The variance for $\hat{\beta}_6$ can be derived similarly, and is given in Table 1. The 'delta method' does not assume that the residuals are normally distributed, so our variance for OLS(Y|X) is more generally applicable than that given in standard textbooks. The expressions for the intercept coefficients are given by $\hat{\alpha}_j = \bar{y} - \hat{\beta}_j \bar{x}$. The expressions for the estimates of the variances of the intercept coefficients associated with $\hat{\beta}_1, \ldots, \hat{\beta}_6$ are given below.

$$\widehat{Var}(\hat{\alpha}_j) = \frac{1}{n^2} \sum_{i=1}^{n} [y_i - \bar{y} - \hat{\beta}_j(x_i - \bar{x}) - \kappa\gamma_{1j}\,(x_i - \bar{x})(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) - \lambda\gamma_{2j}\,(y_i - \bar{y})(y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x}))]^2,$$

where

$$\begin{aligned}
\kappa &= n\bar{x}/S_{xx}, \qquad \lambda = n\bar{x}\hat{\beta}_2/S_{yy}, \\
\gamma_{11} &= \gamma_{22} = 1, \qquad \gamma_{21} = \gamma_{12} = 0, \\
\gamma_{13} &= \hat{\beta}_3\sqrt{(1+\hat{\beta}_2^2)/(1+\hat{\beta}_1^2)}(\hat{\beta}_1 + \hat{\beta}_2)^{-1}, \\
\gamma_{23} &= \hat{\beta}_3\sqrt{(1+\hat{\beta}_1^2)/(1+\hat{\beta}_2^2)}(\hat{\beta}_1 + \hat{\beta}_2)^{-1}, \\
\gamma_{14} &= (\hat{\beta}_4/|\hat{\beta}_1|)(4\hat{\beta}_1^2 + (\hat{\beta}_1\hat{\beta}_2 - 1)^2)^{-1/2}, \\
\gamma_{24} &= \hat{\beta}_4|\hat{\beta}_1|(4\hat{\beta}_1^2 + (\hat{\beta}_1\hat{\beta}_2 - 1)^2)^{-1/2}, \\
\gamma_{15} &= \sqrt{\hat{\beta}_2/4\hat{\beta}_1} \quad \gamma_{25} = \sqrt{\hat{\beta}_1/4\hat{\beta}_2}, \quad \text{and} \quad \gamma_{16} = \gamma_{26} = \tfrac{1}{2}.
\end{aligned}$$

We now proceed to establish several relationships between the six regression slopes and their variances. The discussion is limited to the case where $S_{xy} > 0$, but can easily be extended to $S_{xy} < 0$. The estimates $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$ and $\hat{\beta}_6$ can be written, using simple algebra as:

$$\hat{\beta}_i = a_i + \sqrt{1 + a_i^2}, \quad i = 3, 4, 5 \text{ and } 6, \tag{7}$$

where

$$a_3 = \tfrac{1}{2}(\hat{\beta}_5^2 - 1)\hat{\beta}_6^{-1}, \tag{8}$$

$$a_4 = \tfrac{1}{2}(\hat{\beta}_5^2 - 1)\hat{\beta}_1^{-1}, \tag{9}$$

$$a_5 = \tfrac{1}{2}(\hat{\beta}_5^2 - 1)\hat{\beta}_5^{-1}, \tag{10}$$

$$a_6 = \tfrac{1}{2}(\hat{\beta}_6^2 - 1)\hat{\beta}_6^{-1}. \tag{11}$$

We shall establish the following inequalities for $\hat{\beta}_i$.

Theorem. Suppose $S_{xy} > 0$,

a) If $\hat{\beta}_5 < 1$, then $\hat{\beta}_3 \leq 1$ and

$$\hat{\beta}_1 \leq \hat{\beta}_4 \leq \hat{\beta}_5 \leq \hat{\beta}_3 \leq \hat{\beta}_6 \leq \hat{\beta}_2. \tag{12}$$

b) If $\hat{\beta}_5 \geq 1$, then $\hat{\beta}_3 \geq 1$ and

$$\hat{\beta}_1 \leq \hat{\beta}_3 \leq \hat{\beta}_5 \leq \hat{\beta}_4 \leq \hat{\beta}_2. \tag{13}$$

Further, in this case,

$$\hat{\beta}_4 \leq \hat{\beta}_6 \leq \hat{\beta}_2 \quad if \quad \hat{\beta}_6\hat{\beta}_1 \leq 1, \tag{14}$$

and

$$\hat{\beta}_5 \leq \hat{\beta}_6 \leq \hat{\beta}_4 \quad if \quad \hat{\beta}_6\hat{\beta}_1 > 1. \tag{15}$$

c) If one of $\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$, is equal to 1, then

$$\hat{\beta}_3 = \hat{\beta}_4 = \hat{\beta}_5 = 1. \tag{16}$$

The inequalities are all reversed if $S_{xy} < 0$.

Proof. If $S_{xy} \geq 0$, then clearly

$$0 \leq \hat{\beta}_1 \leq \hat{\beta}_5 \leq \hat{\beta}_6 \leq \hat{\beta}_2. \tag{17}$$

This uses the fact that geometric mean is less than or equal to the arithmetic mean. First note that $\hat{\beta}_1 \leq \hat{\beta}_j \leq \hat{\beta}_2$ for i = 3, 4, 5, 6. Let

$$f(a) = a + \sqrt{1 + a^2}.$$

Clearly $f(a) > 0$ for all a and the derivative of f at a is $f(a)/\sqrt{1 + a^2} > 0$. So

f is a strictly increasing function of a. Further, $f(a)=1$ if and only if $a = 0$. This together with (7)-(10) establishes part (c) of the theorem. If $\hat{\beta}_5 \geq 1$, then by (17),

$$a_3 \leq a_5 \leq a_4 \tag{18}$$

Hence

$$\hat{\beta}_3 \leq \hat{\beta}_5 \leq \hat{\beta}_4. \tag{19}$$

Further note that

$$\begin{aligned} \hat{\beta}_4 &= \tfrac{1}{2}\{\hat{\beta}_2 - \hat{\beta}_1^{-1} + [(\hat{\beta}_2 - \hat{\beta}_1)^2 + 4]^{1/2}\} \\ &= \hat{\beta}_6 - \tfrac{1}{2}(\hat{\beta}_1 + \hat{\beta}_1^{-1}) + \tfrac{1}{2}[(\hat{\beta}_2 - \hat{\beta}_1^{-1})^2 + 4]^{1/2} \end{aligned}$$

Consequently $\hat{\beta}_4 \leq \hat{\beta}_6$ if and only if

$$[(\hat{\beta}_2 - \hat{\beta}_1^{-1})^2 + 4]^{1/2} \leq \hat{\beta}_1 + \hat{\beta}_1^{-1}.$$

This in turn holds if and only if $\hat{\beta}_6\hat{\beta}_1 \leq 1$. This establishes part (b) of the theorem. Similarly if $\hat{\beta}_5 < 1$, then by (8), (9), (10), (11) and (17),

$$a_4 \leq a_5 \leq a_3 \leq a_6 \tag{20}$$

so in this case

$$\hat{\beta}_4 \leq \hat{\beta}_5 \leq \hat{\beta}_3 \leq \hat{\beta}_6 \tag{21}$$

The last part of the theorem follows similarly. This completes the proof of the theorem.

We can establish certain relations among the estimation of variances for the limiting case where $\hat{\beta}_5 = 1$. From Table 1, we have

$$\widehat{Var}(\hat{\beta}_3) = \hat{\beta}_6^{-2}V \tag{22}$$

$$\widehat{Var}(\hat{\beta}_4) = \hat{\beta}_1^{-2}V \tag{23}$$

and

$$\widehat{Var}(\hat{\beta}_5) = V, \tag{24}$$

where

$$V = \frac{1}{4}[\hat{\beta}_2^2 \widehat{Var}(\hat{\beta}_1) + 2\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \hat{\beta}_1^2 \widehat{Var}(\hat{\beta}_2)] \; > \; 0.$$

Clearly from (17), if $\hat{\beta}_5 = 1$, then $\hat{\beta}_1 + \hat{\beta}_2 \geq 2$. So

$$\widehat{Var}(\hat{\beta}_3) \leq \widehat{Var}(\hat{\beta}_5) \leq \widehat{Var}(\hat{\beta}_4) \tag{25}$$

Further from (22) and (24), we have

$$1 \leq \frac{\widehat{Var}(\hat{\beta}_5)}{\widehat{Var}(\hat{\beta}_3)} = \hat{\beta}_6{}^2 = (\frac{1+\hat{\beta}_1^2}{2})^2 \hat{\beta}_1{}^{-2} \rightarrow \infty,$$

as $\hat{\beta}_1 \rightarrow 0$. Finally from (23) and (24),

$$1 \leq \frac{\widehat{Var}(\hat{\beta}_4)}{\widehat{Var}(\hat{\beta}_5)} = \hat{\beta}_1{}^{-2} \rightarrow \infty$$

as $\hat{\beta}_1 \rightarrow 0$.

Though the six $\hat{\beta}_i$ values estimate different quantitative relations between the variables, these inequalities explain the consistent patterns of the $\hat{\beta}_j$ and $\widehat{Var}(\hat{\beta}_j)$ values seen in simulations and real data sets. They may also be useful in applications; for example, a scientist using orthogonal regression would know that this line always lies close to the inverse OLS(X|Y) line when the OLS-bisector slope is $> 45^o$.

SIMULATIONS

We conducted Monte Carlo simulations of data sets with a wide variety of sample sizes, correlation coefficients and linear slopes to test the validity of the formulae given in Table 1. Most were based on bivariate normal samples (σ_x and σ_y are the standard deviations of the distributions in the two variables) with preselected correlation coefficients (ρ). The means μ_x and μ_y are set to 0.0, $\hat{\alpha}$ and $\hat{\sigma}_\alpha$ are the calculated intercepts and their standard deviations, and $\hat{\beta}$ and $\hat{\sigma}_\beta$ are the calculated slopes and their standard deviations.

Table 2 shows typical results for large samples, where we expect the central limit theorem and delta method to be applicable. The table gives the

TABLE II

Linear Regressions for Simulated Bivariate Normal Distributions with Large Samples ($n = 500$)

σ_x	σ_y	ρ	Method	$<\hat{\alpha}>$	$\hat{\sigma}_\alpha$	$<\hat{\beta}>$	$\hat{\sigma}_\beta$	β	$(\hat{\beta}-\beta)/\beta$	$\chi^2(\beta)$
2	1	0.25	OLS(Y|X)	-.00	.04	.125	.022	.125	.00	1.08
			OLS(X|Y)	-.01	.19	2.07	.420	2.00	.04	1.36
			OLS-Bisector	-.00	.07	.709	.039	.708	.00	1.22
			OR	-.00	.04	.162	.028	.162	.00	1.05
			RMA	-.00	.06	.500	.022	.500	.00	1.09
			OLS-mean	-.00	.10	1.10	.199	1.06	.04	1.38
2	1	0.75	OLS(Y|X)	-.00	.03	.375	.015	.375	.00	1.00
			OLS(X|Y)	-.00	.04	.667	.030	.667	.00	0.99
			OLS-Bisector	-.00	.03	.512	.015	.512	.00	0.95
			OR	-.00	.03	.414	.016	.414	.00	0.97
			RMA	-.00	.03	.500	.015	.500	.00	0.95
			OLS-mean	-.00	.03	.521	.016	.521	.00	0.95
1	2	0.25	OLS(Y|X)	.00	.09	.495	.086	.500	-.01	0.99
			OLS(X|Y)	-.00	.37	8.35	1.69	8.00	.04	0.97
			OLS-Bisector	.00	.10	1.41	0.08	1.41	.00	1.03
			OR	.00	.29	6.42	1.30	6.16	.04	0.99
			RMA	.00	.11	2.00	0.09	2.00	.00	1.01
			OLS-mean	.00	.20	4.42	0.81	4.25	.04	0.99
1	2	0.75	OLS(Y|X)	.00	.06	1.50	0.06	1.50	.00	1.05
			OLS(X|Y)	.00	.08	2.68	0.11	2.67	.00	1.07
			OLS-Bisector	.00	.06	1.95	0.06	1.95	.00	1.01
			OR	.00	.07	2.42	0.10	2.41	.00	1.06
			RMA	.00	.06	2.00	0.06	2.00	.00	1.02
			OLS-mean	.00	.07	2.08	0.06	2.08	.00	1.04

mean and standard deviation of the regression coefficients for 500 simulations, each with 500 data points. The final column gives a χ^2 type measure

$$\chi^2(\beta_i) = \frac{1}{M}\sum_{j=1}^{M}(\hat{\beta}_{ij} - \beta_i)^2/\widehat{Var}(\hat{\beta}_i),$$

for β_i given in equations (1)–(6) above, where M denotes the number of simulations and $\hat{\beta}_{ij}$ denotes the estimate of β_i based on j-th simulation. The results indicate that the formulae for all six slopes $\hat{\beta}_i$ are quite accurate for large n. There is less than 1% bias in the average intercept and slope, except for the inverse regression OLS(X|Y) where a slope bias of 2-3% may be present. The $\chi^2(\beta_i)$ values in the Table 2 are close to 1.0, indicating that the estimated variances of regression coefficients reflect the actual dispersion of the slopes. We conclude that our derivations of the coefficients and their variances for the six methods are very accurate for large samples.

Table 2 does reveal, however, considerable differences in the performances of the procedures, as indicated by the dispersions $\hat{\sigma}_\beta$ in the slopes observed from the 500 trials. The inverse OLS(X|Y) consistently has the worst reliability in achieving its theoretical slope. This is due to occasional simulated data sets, where the inverse regression has a nearly vertical slope. Among the methods that treat the variables symmetrically, the OLS-mean and OR lines are generally the least accurate. Unlike OLS-bisector, which is only weakly dependent on the inverse OLS slope, the OLS-mean slope is directly influenced by its linear dependence on the inverse OLS(X|Y) slope. This explains its poor performance.

Table 3 gives analogous results from simulations using small size samples ($n = 50$). When the correlation coefficient is high ($\rho > 0.5$), all six regressions perform well with little bias and reasonably accurate variances ($1.0 \leq \chi^2 \leq 1.3$). The OR slope consistently has the poorest accuracy (*i.e.* largest $\hat{\sigma}_\beta$), the OLS-mean slope has intermediate accuracy, and the OLS-bisector and RMA slopes have the highest accuracy. When the correlation coefficient is

TABLE III

Linear Regressions for Simulated Bivariate Normal Distributions with Small Samples ($n = 50$)

σ_x	σ_y	ρ	Method	$< \hat{\alpha} >$	$\hat{\sigma}_\alpha$	$< \hat{\beta} >$	$\hat{\sigma}_\beta$	β	$(\hat{\beta} - \beta)/\beta$	$\chi^2(\beta)$
2	1	0.25	OLS(Y|X)	-.00	.14	.125	.068	.125	.00	1.00
			OLS(X|Y)	.14	*	3.56	*	2.00	.78	1.94
			OLS-Bisector	.00	.22	.691	.130	.708	-.02	2.35
			OR	-.00	.14	.164	.093	.162	.01	0.99
			RMA	.00	.17	.488	.068	.500	-.02	5.41
			OLS-mean	.07	*	1.84	*	1.06	.74	1.93
2	1	0.75	OLS(Y|X)	.00	.09	.376	.046	.375	.00	1.16
			OLS(X|Y)	.00	.13	.679	.091	.667	.02	1.39
			OLS-Bisector	.00	.10	.517	.047	.512	.01	1.26
			OR	.00	.09	.416	.052	.414	.01	1.17
			RMA	.00	.10	.503	.046	.500	.01	1.23
			OLS-mean	.00	.10	.527	.052	.521	.01	1.26
1	2	0.25	OLS(Y|X)	-.01	.27	.469	.270	.500	-.06	1.22
			OLS(X|Y)	1.01	*	7.33	*	8.00	-.08	1.85
			OLS-Bisector	-.01	.30	1.299	0.24	1.41	-.08	9.52
			OR	.73	*	5.67	*	6.16	-.08	1.95
			RMA	-.00	.35	1.85	0.27	2.00	-.08	11.38
			OLS-mean	.50	*	3.90	*	4.25	-.08	1.82
1	2	0.75	OLS(Y|X)	-.00	.19	1.50	0.19	1.50	.00	1.08
			OLS(X|Y)	-.01	.26	2.74	0.38	2.67	.03	1.26
			OLS-Bisector	-.01	.20	1.96	0.18	1.95	.01	1.24
			OR	-.01	.23	2.47	0.34	2.41	.02	1.27
			RMA	-.01	.20	2.02	0.19	2.00	.01	1.25
			OLS-mean	-.01	.21	2.12	0.21	2.08	.02	1.25

* indicates values greater than 50.

low, the difference between the methods can be dramatic: the scatter in the OLS-bisector and RMA slopes can be many times smaller than the scatter in the OR and OLS-mean slopes. As for the simulations with large-n, the problem with the OR and OLS-mean fits arises when the inverse OLS(X|Y) line becomes nearly vertical and the slope values become very large.

When simulated samples with very small samples ($n < 50$) are considered, the performances of all our regression estimators become progressively less accurate. Particularly when both n and the population correlation ρ are small, our estimators of variances can occasionally be several times too small. Extreme caution should be taken in interpreting regression results under such circumstances.

The small scatter found for the OLS-bisector slope in these simulations shows that the analytic result of equation (25), which strictly applies only under limited circumstances, empirically applies over a wide range of circumstances. From these simulations we conclude that, if precision in the regression coefficients is the sole criterion for preferring one method over another, the standard OLS(Y|X), OLS-bisector and RMA lines are the best performers (and generally have similar scatter in their coefficients), while the inverse OLS(X|Y), OR and OLS-mean lines perform relatively poorly.

CONCLUSIONS

The work described here establishes that the regression slope and variance estimates given in Table 1 are accurate for moderate to large samples: the slope estimates reproduce the theoretical values for normal models without bias, and the variance estimates are consistent with empirical measures of slope dispersion. We also establish analytical relations between the various slopes and variances. They show that, under specified broad conditions, certain slopes are always steeper/flatter than others.

Our study of the relations among these six formally unrelated fits can elucidate disagreements among scientific researchers using different methods

to address the same problem. It should be useful, for example, in clarifying astronomers' efforts to quantify the 'cosmic distance scale'. Detailed practical guidelines derived from this study are presented in the Conclusion section of IFAB, and are summarized here. We observe that the standard OLS(Y|X) performs well and should be preferred when the distinction between dependent and independent variable is clear. Among the methods treating the variables symmetrically, the OLS-bisector and reduced major axis have smaller variances than Pearson's orthogonal regression and the arithmetic-mean OLS line. But, the reduced major axis has a crucial limitation (see IFAB and references therein): its amplitude depends only on the population dispersions and is independent of the population correlation ρ (except for its sign, see equation 5). Though a least-squares procedure, it does not measure any relation between the variables. We thus conclude that the OLS-bisector may be the most desirable regression line symmetric in the two variables. It has a small variance, and does depend (though in a complicated fashion; see equation 3) on the population correlation.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Marllyn Boswell (Penn State) for assistance with the numerical simulations, and Drs. Takashi Isobe (MIT) and Michael Akritas (Penn State) for their advice and assistance. Astronomical statistics at Penn State is sponsored by NASA grants NAGW-1917 and NAGW-2120, NSF grant DMS-9007717, and the Center of Excellence in Space Data and Information Sciences (operated by the Universities Space Research Association in cooperation with NASA). The authors would also like to thank the referees for helpful suggestions, which improved the presentation.

BIBLIOGRAPHY

Aaronson M., Bothun, G., Moutld J., Huchra, J., Schommer, R. A., and Cornell, M. E. (1986), 'A Distance Scale from the Infrared Magnitude/H I Velocity-Width Relation. V. Distance Moduli to 10 Galaxy Clusters, and

Positive Detection of Bulk Supercluster Motion Toward the Microwave Anisotropy', *Astrophysical Journal, 302,* 536-563.

Bevington, P. R. (1969), *Data Reduction and Error Analysis for the Physical Sciences*, New York:McGraw-Hill.

Billingsley, R. (1986), *Probability and Measure* (2nd Ed.), New York: John Wiley.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton: Princeton University Press.

Freedman, W. L. (1990), 'Local Calibrators for the Infrared Tully-Fisher Relation', *Astrophysical Journal Letters, 355,* L35-L38.

Isobe, T., Feigelson, E. D., Akritas, M. G. and Babu, G. J. (1990), 'Linear Regression in Astronomy. I.', *Astrophysical Journal, 364,* 104-113.

Jungers, W . L. (editor) (1985), *Size and Scaling in Primate Biology*, New York:Plenum.

Kermack, K. A. and Haldane, J. B. S. (1950), 'Organic Correlation and Allometry', *Biometrika, 37,* 30-41.

Krutchkoff, R. G. (1967), 'Classical and Inverse Regression Methods of Calibration', *Technometrics, 9,* 425-439.

Linnik, Yu. V. (1961), *Method of Least Squares and Principles of the Theory of Observations*, New York:Pergamon.

Massart, D. L., Dijkstra, A., and Kaufman, L. (1978) *Evaluation and Optimization of Laboratory Methods and Analytical Procedures*, Amsterdam:Elsevier.

Pearson, K. (1901), 'On Lines and Planes of Closest Fit to Systems of Points in Space', *Philosophical Magazine, Series 6, 2,* 559-572.

Pierce, M. J. and Tully, R. B. (1988) 'Distances to the Virgo and Ursa Major Clusters and a Determination of H_o, *Astrophysical Journal, 330,* 579-595.

Rowan-Robinson, M. (1985), *The Cosmological Distance Ladder*, New York:W. H. Freeman.

Rubin, V. C., Burstein, D. and Thonnard, N. (1980), 'A New Relation for Estimating the Intrinsic Luminosities of Spiral Galaxies', *Astrophysical Journal Letters, 242,* L149-L152.

Rubin, V. C. and Coyne, G. V. (Eds.) (1989), *Large-Scale Motions in the Universe*, Princeton: Princeton University Press.

Shorter, J. (1982), *Correlation Analysis of Organic Reactivity*, Chicester:Res. Studies Press.

Sokal, R. R. and Rohlf, F. J. (1981) *Biometry:The Principle and Practice of Statistics in Biological Research*, 2nd ed., San Francisco:W. H. Freeman.

Strömberg, G. (1940), 'Accidental and Systematic Errors in Spectroscopic Absolute Magnitudes for Dwarf G0-K2 Stars', *Astrophysical Journal, 92,* 156-169.

Tully, R. B., (1988) 'Origin of the Hubble Constant Controversy', *Nature, 334,* 209-212.

Received April 1991; Revised August 1991