THE ASTROPHYSICAL JOURNAL, 397:55–67, 1992 September 20


# LINEAR REGRESSION IN ASTRONOMY. II.

ERIC D. FEIGELSON
Department of Astronomy and Astrophysics, Pennsylvania State University, Davey Laboratory, University Park, PA 16802
AND
GUTTI JOGESH BABU
Department of Statistics, Pennsylvania State University, Pond Laboratory, University Park, PA 16802
*Received 1991 December 13; accepted 1992 March 19*

## ABSTRACT

A wide variety of least-squares linear regression procedures used in observational astronomy, particularly investigations of the cosmic distance scale, are presented and discussed. We emphasize that different regression procedures represent intrinsically different functionalities of the data set under consideration and should be used only under specific conditions. Discussion is restricted to least-squares approaches, and for most methods computer codes are located or provided. The classes of linear models considered are (1) unweighted regression lines, some discussed earlier in Paper I of this series, with bootstrap and jackknife resampling; (2) regression solutions when measurement error, in one or both variables, dominates the scatter; (3) methods to apply a calibration line to new data; (4) truncated regression models, which apply to flux-limited data sets; and (4) censored regression models, which apply when nondetections are present.

For the calibration problem we develop two new procedures: a formula for the intercept offset between two parallel data sets, which propagates slope errors from one regression to the other; and a generalization of the Working-Hotelling confidence bands to nonstandard least-squares lines. They can provide improved error analysis for Faber-Jackson, Tully-Fisher, and similar cosmic distance scale relations. We apply them to a recently published data set, showing that the distance ratio between the Coma and Virgo clusters can be determined to $\sim 1\%$ accuracy.

The paper concludes with suggested strategies for the astronomer in dealing with linear regression problems. Precise formulation of the scientific question and scrutiny of the sources of scatter are crucial for optimal statistical treatment.

*Subject headings:* distance scale — methods: numerical

## 1. INTRODUCTION

Linear regression is a fundamental and frequently used statistical tool in all branches of observational astronomy. It is used in exploratory data analysis to quantify trends, in observational tests of astrophysical theory, and in the cosmic distance scale. In distance scale applications, especially, accurate regression coefficients and error analysis are crucial to measuring the expansion rate of the universe, estimating the age of the universe, and uncovering large-scale phenomena such as superclustering and galaxy streaming. It is perhaps initially surprising that an apparently simple statistical procedure, like least-squares linear regression with two variables, is difficult or controversial. Yet recent cosmic distance scale studies have utilized at least eight different methods for similar tasks.

The complexity in linear regression analysis arises for a variety of reasons: properties of the data are not always the same (e.g., the scatter may or may not have a Gaussian distribution, the data may be heteroscedastic where the degree of scatter depends on the data value in one or both variables); scientists' knowledge of the situation is not always the same (e.g., they may or may not know how much of the scatter in each variable is due to the measurement processes rather than to the objects under study); and the purpose of the analysis is not always the same (e.g., one problem might need an optimized estimate of the intercept, and another problem may seek to apply the regression to new data points). Statistical research into linear regression methods dependent on all of these issues started in the nineteenth century and is still very active.

We concentrate consideration here on least-squares linear regression problems that, in our judgment, are particularly relevant to applications in astronomy, such as the cosmic distance scale.[1] These are the following: choice of regression line without weighting (§ 2); applying a calibration regression to new data (§ 3); weighted regression with known measurement errors (§ 4); regression with truncation (§ 5); and regression with censoring (§ 6). Section 7 gives sample applications of methods we develop to astronomical data sets. Finally, conclusions and recommendations are summarized in § 8. The subject of § 2 is covered only briefly here, as it is discussed in detail in our previous papers (Isobe et al. 1990 and Babu & Feigelson 1992, hereafter IFAB and BF, respectively). For some topics (§§ 5–7) we provide mainly reviews of the extensive literature, practical information for the astronomer. For other problems (§§ 2 and 3) we develop new statistical methodology (Appendix). Diagrams illustrating the six situations discussed in §§ 2–7 are shown in Figures 1–6; the diagrams thus provide a visual guide to the paper.

[1] To maintain a reasonable scope, we have purposely omitted from this study a wide variety of non–least-squares linear regression methods. These include "robust" methods which minimize the absolute deviations rather than the squared deviations of points from the line (see Branham 1982, Lutz 1983, and Press et al. 1986, chap. 14.6, for discussions in astronomical contexts), standard multivariate regression methods, ridge and related Bayesian regressions for multivariate problems with collinearities (Hocking 1983; Amemiya 1985, chap. 2), and slope estimates based on nonparametric statistics (Sen 1968).

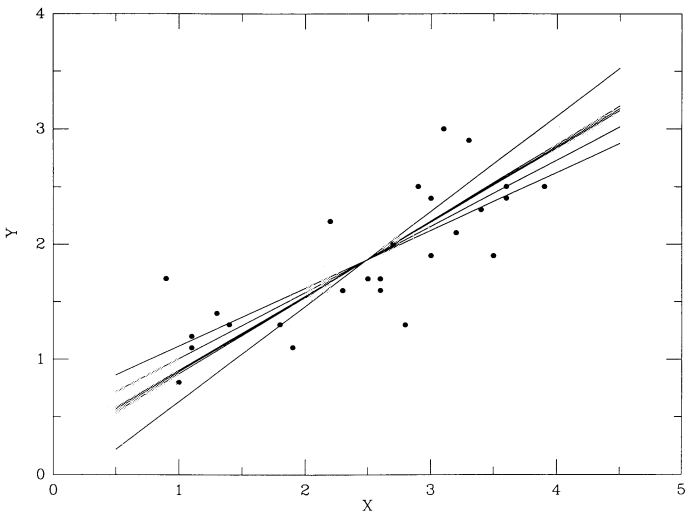


FIG. 1.—[Figs. 1–6 are diagrams of the bivariate regressions discussed in the text, using a hypothetical data set with 26 points.] Unweighted least-squares models (§ 2). Starting with the steepest line, they are OLS($X \mid Y$), OLS mean, OLS bisector, reduced major axis, orthogonal, and OLS($Y \mid X$) regressions. The lines are calculated using the program SLOPES.

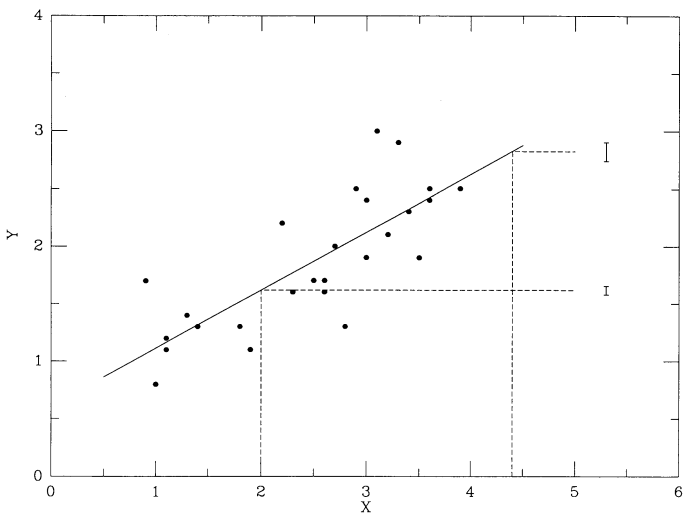


FIG. 3.—Applying a calibration line to new data points (§ 3.2). The predicted $Y$-values and uncertainties are calculated using program CALIB.

## 2. UNWEIGHTED LINEAR REGRESSION MODELS

### 2.1. ($x_i$, $y_i$) *Data Sets with No Additional Information*

In a great many astronomical studies, one seeks a linear fit to data sets of the form ($x_i$, $y_i$), where no additional information is available. Measurement error is relatively unimportant, and the degree of scatter in each variable is unknown. This regression problem was addressed in detail in IFAB and BF, and we review the findings here. Astronomers have used as many as six different unweighted least-squares linear regression lines for this situation: the standard ordinary least-squares solution OLS($Y \mid X$), which minimizes the residuals in $Y$; the solution OLS($X \mid Y$), which minimizes the residuals in $X$; the orthogonal regression line, which minimizes the perpendicular distances; the OLS bisector, the line which bisects the OLS lines; the reduced major axis, the geometric mean of the OLS lines; and the OLS mean, the arithmetic mean of the OLS lines. The last four lines are invariant to a switch of variables, and were developed for problems where the choice of "dependent" and "independent" variables is unclear. We show (e.g., IFAB eqs. [22]–[26]) that these six solutions calculate intrinsically different functionalities of the data. Scientific results, such as galaxy distances or the Hubble constant, obtained with different lines are not directly comparable, any more than results using linear and exponential fits can be compared. No line is theoretically "better" than another when there is no prior information (including clear specification of the goal of the regression) other than the ($x_i$, $y_i$) values under study.

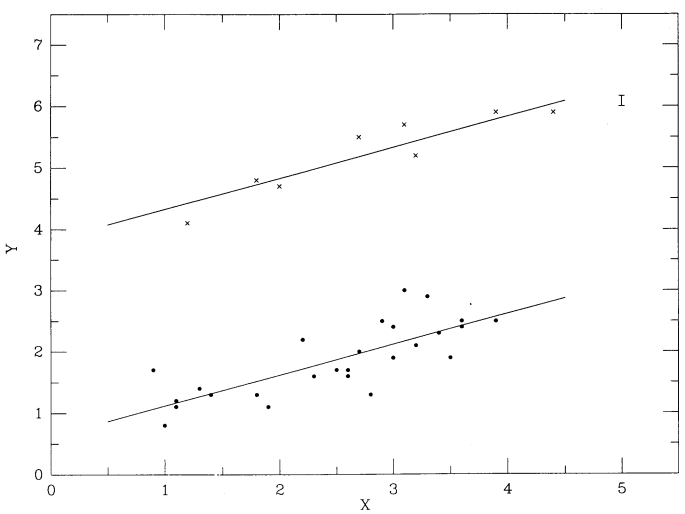


FIG. 2.—Applying a calibration line to a new data set, to find the intercept offset (§ 3.1). The intercept and its uncertainty (*error bar at upper right*) are calculated using eqs. (A3) and (A4), using OLS($Y \mid X$) of the lower data set as the calibration relation.

While this indefiniteness is disquieting, some guidance regarding choice of line can be provided. If it is known independently of the data set under consideration that one variable physically depends on another, or if the scientific question clearly asks how one variable depends on the other, then there is a preference to use OLS($Y \mid X$) with $Y$ as the dependent variable. The reduced major axis (known to astronomers as Strömberg's impartial line; Strömberg 1940) is discouraged, since it does not scale with the population correlation coefficient. The orthogonal regression line is applicable only to scale-free (e.g., log) variables. The inverse solution OLS($X \mid Y$) is particularly unstable for small samples and/or low correlation coefficients. We conclude that two of the lines—the standard OLS($Y \mid X$) and the OLS bisector—in practice achieve their theoretical regression coefficients more accurately than the other lines, and are therefore recommended for use.

Furthermore, it is clearly established that use of the usual standard formulae for the uncertainties of the derived slopes and intercepts (e.g., Bevington 1969) is mathematically incorrect except for OLS($Y \mid X$) under restrictive assumptions. Table 1 in IFAB and BF provide correct asymptotic (i.e., for large $N$) error formulae for each of the six unweighted regression lines. However, these asymptotic formulae underestimate the true regression coefficient uncertainty in samples with small $N$ (approximately $N \leq 50$) or small population correlation. Resampling procedures such as the jackknife or bootstrap

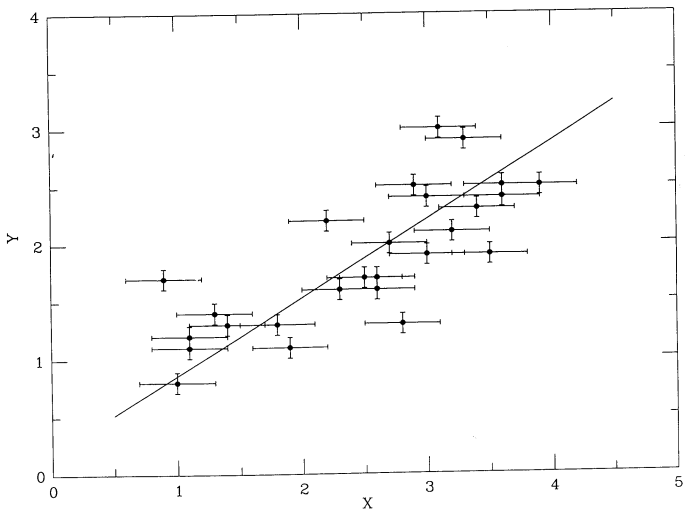


FIG. 4a

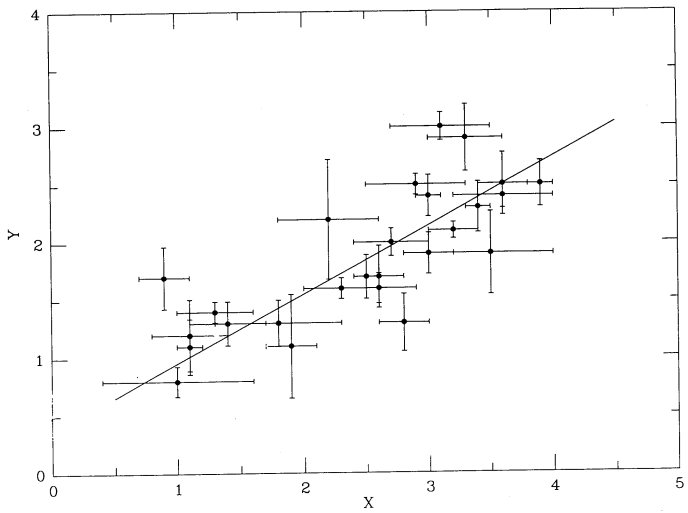


FIG. 4b

FIG. 4.—(*a*) Homoscedastic functional measurement error model (§ 4.1), shown here with $\lambda = \frac{1}{3}$. The line is the solution to eq. (1). (*b*) Heteroscedastic functional measurement error model (§ 4.2). The line is the solution to eq. (11) calculated using the algorithm of Ripley & Thompson (1987).

should be used to evaluate regression uncertainties in these cases.[2]

A short FORTRAN subroutine, SIXLIN, has been distributed that implements these six regression lines and the asymptotic error analysis (IFAB). We announce here the availability of a FORTRAN program called SLOPES that runs SIXLIN, and also calculates error analyses based on bootstrap and jackknife resampling, and bivariate normal numerical simulation, of the data set, as discussed in BF. Bootstrap error analysis is based on the distribution of slopes and intercepts of a large number of data sets constructed by random sampling of the observed data set with replacement. Thus, each bootstrapped data set has $N$ points drawn from the original sample, where some original points are present two or more times and other original points are missing. Jackknife error analysis is performed in a similar way but with only $N$ synthetic data sets, each containing $(N - 1)$ points from the original data set, leaving out one observation in sequence. The use of bootstrap and jackknife techniques to estimate confidence intervals for small or non-normal data sets is discussed by Efron & Tibshirani (1986), Babu (1992), and references therein. The simulation of bivariate normal data sets with properties matching the observed data set are calculated using the program TULSIM (Boswell 1990). SLOPES can be obtained without charge from the authors (Internet: code@stat.psu.edu).

[2] Two equations in Appendix A of IFAB contain typographical errors: the first occurrences of the factor $\beta_1$ on the right-hand sides of eqs. (A3) and (A6) should be $\beta_1^{-1}$. The covariance definition at the bottom of Table 1 and the implementation in the FORTRAN code SIXLIN are correct. We are grateful to V. M. Blanco for pointing out this error.

### 2.2. ($x_i$, $y_i$) *Data Sets with Information on the Scatter*

In a few situations the astronomer may know, independently of the sample at hand, the scatter of the intrinsic properties

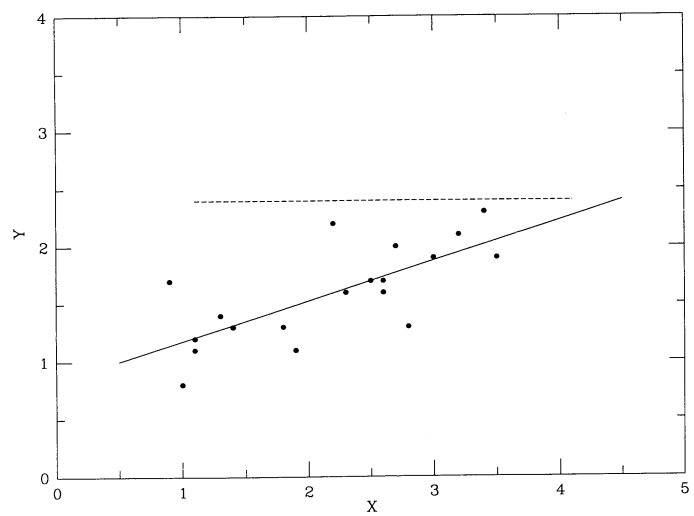


FIG. 5.—Truncated regression model (§ 5), where the $Y$ variable is assumed to be truncated just at 2.4, so six points from the sample are missing. The line is calculated using LIMDEP (Greene 1989).

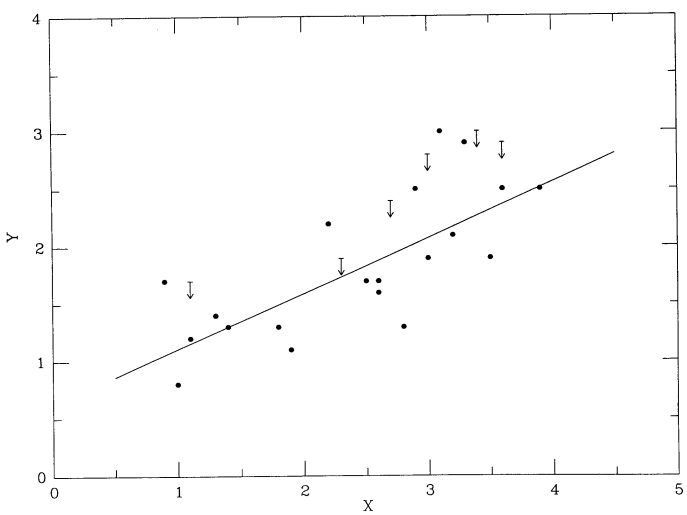


FIG. 6.—Censored regression model (§ 6), where several $Y$-values are assumed to be undetected at various sensitivity levels. The line is calculated using ASURV.

under study. For example, the width of the main sequence in globular cluster color-magnitude diagrams may serve as a measure of the scatter of stellar luminosities, the scatter of a variable in a well-studied calibration data set might be applied to a new sample, or repeated observations of the same objects might be studied. The regression solution for one such class of problems is well established. If the intrinsic variance in $X$ is $\sigma_x^2$ and in $Y$ is $\sigma_y^2$, one can define the ratio $l = \sigma_x^2/\sigma_y^2$. The solutions for $l = 0$, 1, and $\infty$ are simply the OLS($X \mid Y$), orthogonal, and OLS($Y \mid X$) lines discussed in § 2.1. For arbitrary $l$ and assuming all residuals are distributed as Gaussians, the least-squares estimate of the slope is

$$\hat{\beta} = \{S_{yy} - lS_{xx} + [(S_{yy} - lS_{xx})^2 + 4lS_{xy}^2]^{1/2}\}/(2S_{xy}) , \quad (1)$$

where

$$S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2 , \quad S_{yy} = \sum_{i=1}^{N} (y_i - \bar{y})^2 ,$$

$$S_{xy} = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) , \quad (2)$$

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

The corresponding uncertainty of this slope is

$$\sigma_{\beta} = [\widehat{\mathrm{Var}}\ (\hat{\beta})]^{1/2} , \quad (3)$$

where

$$\widehat{\mathrm{Var}}\ (\hat{\beta}) = \frac{\hat{\beta}^2}{S_{xy}^2}\left[\frac{S_{xy}R}{\hat{\beta}} + \frac{(S_{yy} - \hat{\beta}S_{xy})R}{l} - \frac{\hat{\beta}^2}{(n-1)l^2}(S_{yy} - \hat{\beta}S_{xy})^2\right] \quad (4)$$

and

$$R = \frac{1}{n-2}\sum_{i=1}^{n} [y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})]^2 . \quad (5)$$

For the more general case where the residuals are not assumed to have Gaussian distributions, the appropriate slope variance is

$$\widehat{\mathrm{Var}}\ (\hat{\beta}) = \frac{\hat{\beta}^2}{(\hat{\beta}_2^{-1} - l\hat{\beta}_1)^2 + 4l^2} \times [l^2\ \widehat{\mathrm{Var}}\ (\hat{\beta}_1) + \hat{\beta}_2^{-4}\ \widehat{\mathrm{Var}}\ (\hat{\beta}_2) + 2l\hat{\beta}_2^{-2}\ \widehat{\mathrm{Cov}}\ (\hat{\beta}_1, \hat{\beta}_2)] , \quad (6)$$

where $\widehat{\mathrm{Var}}\ (\hat{\beta}_1)$, $\widehat{\mathrm{Var}}\ (\hat{\beta}_2)$, and $\widehat{\mathrm{Cov}}\ (\hat{\beta}_1, \hat{\beta}_2)$ are given in Table 1 of IFAB. The derivation of this variance is similar to that of orthogonal regression (IFAB, Appendix A) using the $\delta$-method. In all cases, the intercept is determined from $\alpha = \bar{y} - \hat{\beta}\bar{x}$.

Note that these models should only be used with scale-free (e.g., log) variables (except when $l = 0$ or $\infty$). As with the orthogonal regression line, these slopes will change when one of the variables is expressed in different units.

### 2.3. *Slopes Estimates Based on* $\overline{y_i/x_i}$

Astronomers have occasionally evaluated slopes of unweighted bivariate relations using the simple average of the ratios $y_i/x_i$. This procedure has arisen in the estimation of Hubble's constant from samples of galaxies in rich clusters, where $H_0$ is estimated to be the mean of individual galaxy Hubble ratios $v_i/r_i$ within a given rich cluster (e.g., Aaronson et al. 1986; Pierce & Tully 1988; Fouqué et al. 1990). While this does not at first appear to be a regression procedure at all, it is in fact a least-squares solution to the heteroscedastic model:

$$y_i = \beta x_i + \eta_i x_i , \quad (7)$$

where the intercept is zero and $\eta_i$ is the random error in $Y$ with an expected value of zero. Here the scatter in $Y$ is no longer independent of $X$, but increases linearly with $X$. The conditions underlying this model, which is studied in econometrics and other fields, rarely arise in astronomy. It requires the intercept to be precisely zero, the $x_i$ values to be known without error, and the scatter in $Y$ to increase linearly with $X$. In the particular case of the Hubble diagram, the intercept is difficult to remove due to Virgo infall and other local galaxy motions; the distances are most certainly subject to measurement error due to imperfect distance indicators; and the principal source of error in these velocities is probably from non-Hubble motions which do not scale with distance. We therefore discourage the use of averages of $y_i/x_i$ ratios for $H_0$ estimates and most other astronomical applications. $H_0$ should instead be estimated by regressions on the $(r_i, v_i)$ Hubble diagram.

## 3. CALIBRATION

Many cosmic distance scale calculations involve a sequence of two or more linear regressions. A regression is first calculated for a small sample of nearby calibrator objects with estimable distances (e.g., galaxies sufficiently close for individual Cepheid variables to be studied, stars sufficiently close for annual parallax measurement). This calibration line is then applied to more distant objects to estimate their distances.

Statistically, calibration "is the process of assigning values to the response of an instrument [e.g., photometer or spectrograph] or the property of an artifact [e.g., star or galaxy] relative to reference standards or measuring processes," where the reference standard is measured with negligible error (Cameron 1982). This meaning is now called "absolute calibration." Astronomers perform absolute calibration when an astronomical spectrum is compared with the emission-line wavelengths of a Th-Ar lamp. For the univariate linear case, the standard OLS($Y \mid X$) was usually used, with careful attention paid to error propagation and the range over which the calibration is accurate (see, for example, Scheffé 1973; Kendall & Stuart 1977, pp. 385–393; Neter, Wasserman, & Kutner 1985, chap. 5; Anderson 1987, chap. 6). A vigorous debate has emerged over the relative merits of OLS($Y \mid X$) and the inverse OLS($X \mid Y$) for problems where the $Y$ variable of a new object is measured to predict its $X$-value from the calibration line (Krutchkoff 1967; see Martens & Naes 1989, pp. 73–85, and Osborne 1991, § 2, for reviews of the debate). A comprehensive review of statistical calibration issues (including inverse, Bayesian, nonlinear, nonparametric, multivariate, and other models) is given by Osborne.

The astronomical problems we address here, such as calibrations of cosmic distance scale indicators, are examples of "comparative" calibration. Here "one instrument or measurement technique is calibrated against another, with neither one being inherently a standard" (Rosenblatt & Spiegelman 1981). Astronomical calibration samples are usually random stars or galaxies which happen to lie sufficiently near us that certain important properties can be directly measured. These calibration samples can suffer both nonnegligible measurement

errors and intrinsic scatter about the line, and thus are not absolute standards like emission lines from a Th-Ar lamp. The mathematical tools used for this type of calibration must take these effects into account.

We provide here two procedures for evaluating the uncertainties of values for the new sample after application of a linear relation derived from the calibration sample. The first (§ 3.1) calculates the offset in intercepts between two parallel data sets. The second procedure (§ 3.2) predicts $Y$-values for the new objects individually, propagating errors fully. Our procedures apply to any of the six unweighted regression lines described in § 2; we thus sidestep the "classical" versus "inverse" calibration debate, allowing scientists to compare calibrations using different least-squares lines. We assume that the values of the $Y$ variable of the new sample are to be estimated, given the $X$- and $Y$-values of the calibration sample and the $X$-values of the new sample. We also assume that the new and calibration samples both follow a linear relation with the same slope.

### 3.1. *Applying a Calibration Line to a New Data Set*

In a few cosmic distance scale applications, such as estimating the difference in distances between clusters of galaxies, only the intercept of the new sample, rather than the distance to each individual new galaxy, is desired. The statistical problem here is to find the optimum intercept offset, and its uncertainty, between the calibrator and the new regression lines. Various solutions have been proposed for this problem: standard parametric comparison of regression lines (e.g., Neter et al. 1985, chap. 3); likelihood ratio tests for equality of intercepts over restricted ranges of $X$ (Spurrier, Hewett, & Lababidi 1982); nonparametric and parametric estimation of intercept differences based on Sen's (1969) parallelism test (Akritas, Saleh, & Sen 1985; Lambert, Saleh, & Sen 1985); and an $F$-estimator for the intercepts (Saleh & Hassanein 1986). These past treatments, however, apply only under restricted conditions (e.g., $X$-values have no error, $Y$ residuals have finite Fisher information).

We address this problem in the context of the six regression lines described in § 2. In the spirit of the regression coefficient variance derivations in IFAB, which make no assumption regarding the underlying distributions and permit both the $X$ and $Y$ variables to be random, we derive in § A1 of the Appendix the intercept offset uncertainty for a calibration least-squares regression line applied to a new data set. Equations (A4) and (A6) show that the new intercept uncertainty depends on a term relating to the vertical scatter in the new sample, and terms propagating the uncertainties of the slope and intercept of the calibrator relation.

### 3.2. *Applying a Calibration Line to a New Data Point*

The most commonly accepted solution to calculating the uncertainty of values obtained from the calibration line was derived by Working & Hotelling (1929).[3] They showed that, when given a new $x_i$ and seeking its $y_i$ value from the calibration regression, the joint confidence bands (curves of constant confidence level above and below the calibration line) are hyperbolae. This is easily understood: the narrow confidence region in the middle of the distribution is due to the uncertainty in intercept, and the wider confidence region at the ends of the distribution is due to the uncertainty in slope. The standard Working-Hotelling confidence bands, however, have two limitations for astronomical calibration problems: they are based on the assumption that the $X$ variable is experimentally controlled, whereas in astronomy it is usually a random uncontrolled variable; and they are given only for OLS($Y \mid X$), whereas the astronomer might desire confidence bands for the OLS bisector or other OLS line.

We derive in § A2 of the Appendix asymptotic (i.e., valid for large $N$) formulae giving Working-Hotelling–type confidence bands without these restrictions. Points at the ends of the distribution will have less accurate predicted values than those near the middle. A FORTRAN code entitled CALIB is available from the authors (Internet: code@stat.psu.edu) which computes the six OLS lines described in § 2 for a calibrator sample and applies them to a sample of new $x_i$ values to give predicted $y_i$ and $\sigma(y_i)$ values. If the calibration relation has small $N$ ($<50$), we recommend that users first run the program SLOPES (see § 2) on the calibration sample to see whether the regression coefficient uncertainties derived from bootstrap and jackknife simulations are significantly larger than the asymptotic uncertainties. If so, then our asymptotic confidence band formulae probably also underestimate the true uncertainties $\sigma(y_i)$ by a similar factor.

## 4. MEASUREMENT ERROR REGRESSION MODELS

In some astronomical regressions, the scatter about the line is due primarily to the measurement process. In statistics, problems where the true points lie precisely on the line are called "functional" regression models, while problems where the true points are scattered about the line (such as those discussed in the previous sections) are called "structural" regression models. Functional models are appropriate when measurement errors exceed intrinsic scatters, or when the intrinsic scatter is obviously zero (e.g., measurements of the same property and objects using different detectors).

The literature on such regression models is enormous: the methods are variously called weighted regression, "errors-in-variable" models, measurement error models, path analysis, and latent or instrumental variable models. Extensive bibliographies and theoretical treatments can be found in Fuller (1987), Madansky (1959), and Anderson (1984). Less mathematical reviews that we recommend for astronomers included Jones's (1979) treatment for geologists, Mandel's (1984) for engineers, Judge et al.'s (1988, chap. 19) for economists, and Deeming's (1968) long-neglected paper for astronomers. Other interesting perspectives on measurement error models include a Monte Carlo comparison of more than 30 functional regression models by Riggs, Guarnieri, & Addelman (1978); a comparison of several measurement error computer codes accessible to astronomers by Murtagh (1990); and a Bayesian analysis by Reilly & Patino-Leal (1981).

### 4.1. *Homoscedastic Functional Model*

In some astronomical problems the scatter due to errors in the measurement process is the same from point to point. This condition is called homoscedasticity. An astronomical example of this might be an optical color-magnitude diagram where $V$-magnitudes are subject to $\pm 0.05$ mag and $B-V$ are subject to $\pm 0.03$ mag uncertainties. This is the classical "errors-in-variables" regression model, and has been extensively dis-

[3] This is not the only approach. See, for example, Neter et al. (1985, chap. 5) for a comparison of Working-Hotelling, Bonferroni, Gafarian, and Scheffé joint confidence intervals, and Hunter & Lamboy (1981) for a Bayesian approach.

cussed since the 1870s. The model is usually expressed as follows: if the true relation is

$$Y = \alpha + \beta X , \quad (8)$$

the measured quantities are

$$x_i = X_i + \delta_i , \quad y_i = Y_i + \epsilon_i , \quad (9)$$

and the ratio of measurement error variances is defined to be

$$\lambda_i = \sigma^2(\epsilon_i)/\sigma^2(\delta_i) . \quad (10)$$

Many of the published methods apply only to restricted situations. For instance, when $\lambda_i = l$ = constant, the least-squares estimate of the slope is given by equation (1), with variance estimate given by equation (4). A thorough treatment of this model is given by Fuller (1987, pp. 30–50). This variance is applicable only in cases of large samples and normal residuals. The exact distribution of the slope is quite complicated (Anderson & Sawa 1982), and reliable error analysis for the small-sample or nonnormal structural model often requires resampling techniques like the bootstrap. This solution is only one of a class of homoscedastic functional measurement error models with analytical solutions; see Deeming (1968) and Jones (1979) for clear presentations of the field.

Software systems EV-CARP and SUPER-CARP implement many analytical solutions from "errors-in-variables" statistics and can be purchased from Fuller (1988).

### 4.2. *Heteroscedastic Functional Model*

For the more general heteroscedastic case, where the measurement errors are different for each point and each variable ($\lambda_i \neq$ constant) but the intrinsic points are still assumed to lie exactly on a line, weighted regressions are used. If the measurement error is confined to the dependent variable, then weighting each by $1/\sigma^2(\delta_i)$ and applying the OLS($Y \mid X$) line is standard (e.g., Bevington 1969). When $\lambda_i \neq 0$ and heteroscedastic measurement errors are present in both variables, then a "doubly weighted" regression is the appropriate generalization of OLS($Y \mid X$). This situation occurs when different objects are observed with different signal-to-noise ratios, or when the luminosities rather than fluxes are considered and the objects lie at different distances. Here one finds regression coefficients $\alpha$ and $\beta$ that minimize the quantity

$$S = \sum [(x_i - \hat{x}_i)^2/\sigma^2(\delta_i) + (y_i - \hat{y}_i)^2/\sigma^2(\epsilon_i)] , \quad (11)$$

where $(\hat{x}_i, \hat{y}_i)$ are the observed $(x_i, y_i)$ "adjusted" to lie on the $Y = \alpha + \beta X$ line by sliding them along a line of slope $\lambda_i$ (see York 1967 for a pictorial description of this process). The line satisfying this condition can be found analytically, as a root of a cubic equation (York 1966; McIntyre et al. 1966) or by iterative numerical calculation. A difficult manual iterative procedure was described by Deming (1943, pp. 178 ff.), and various computerized approaches have been widely discussed in the physics and statistics literature (Boggs, Byrd, & Schnabel 1987 and references therein).

Software implementing the doubly weighted regression model include GaussFit (Jefferys et al. 1988a, b), Fasano & Vio (1988), Ripley & Thompson (1987), and ODRPACK (Boggs et al. 1990). Several implementations of these numerical methods are compared using astronomical data sets by Murtagh (1990); all give identical slope estimates, but their error estimates may differ. ODRPACK (standing for "orthogonal distance regression" package), which treats nonlinear as well as linear fits, is available in the public domain (ftp research.att.com, username "netlib").

### 4.3. *Measurement Error Structural Models*

Many astronomical problems will possess scatter both from the intrinsic properties of the objects and from the measurement process. One well-developed approach to such mixed structural-measurement error models is LISREL (Jöreskog 1973), which has been extensively used in social science applications. LISREL is both mathematically and computationally complex, and a software implementation is widely available (LISREL VI; Jöreskog & Sörbon 1984). However, the LISREL model is restricted to multivariate problems (three or more independent variables) with homoscedasticity ($\lambda_i = \lambda =$ constant) and Gaussian residuals. It may be appropriate for astronomical problems satisfying these constraints.

## 5. TRUNCATED REGRESSION MODELS

Any of the regression lines listed above will clearly be biased if one or both variables are subject to truncation—that is, when any points above or below some value are omitted from the sample. Although a general solution to recovering the true line from truncation bias in either variable for any of the six OLS lines has not been achieved, solutions for OLS($Y \mid X$) with truncation in $Y$ have been obtained. This problem has been investigated in econometrics, where, as in astronomy, samples frequently suffer a selection bias when objects fall above or below a known truncation limit. Variables with such biases are known as "limited-dependent" variables, and maximum-likelihood solutions assuming Gaussian distributions are often treated under the rubric of the "Tobit" regression models (see Maddala 1983, chaps. 6 and 7; Amemiya 1985, chap. 10; and Greene 1991, chap. 21). Truncated regression has also been discussed in astronomical contexts by Segal (1975) and Teerikorpi (1984); regression coefficients are not directly calculated in the latter study.

Several semiparametric approaches to the truncated linear regression problem, which do not assume normal distributions in the residuals, have been recently studied. An estimate of the regression slope based on a Kolmogorov-Smirnov test on sequential ranks of the data in both variables was proposed by astronomer Turner (1979), and its asymptotic validity was established by statistician Bhattacharya (1983). The efficiency of this test, however, is unknown. Another slope estimate based on Kendall's $\tau$ was derived by Bhattacharya, Chernoff, & Yang (1983). The corresponding confidence intervals are described by Antille & Milasevic (1988) when replicate observations at each value of $X$ are available. Here contributions of points near the truncation limit are iteratively removed from the calculation, which may eliminate a significant fraction of the data. A third semiparametric method iteratively adjusts dependent values near the truncation level (Tsui, Jewell, & Wu 1988). If the residuals are Gaussian, the model becomes the maximum-likelihood Tobit solution. Confidence intervals for the derived slopes are not known, though they may be estimatable using bootstrap simulations. Other recent methodological approaches probably relevant to astronomical flux-limited data sets are those of Vardi (1985), Bickel & Ritov (1991), and Lai & Ying (1992).

The principal import for astronomy (particularly cosmic distance scale problems based on flux-limited samples) is that modifications to OLS($Y \mid X$) that compensate for the truncation in $Y$ are known, and should be used when truncation bias

may be important. Since research on semiparametric models is still ongoing, we suggest that astronomers adopt the maximum-likelihood Tobit solutions from econometrics (which assume normal residuals). We discourage—in contrast to a number of astronomical studies (see Tully 1988 and references therein)—use of the inverse solution OLS($X \mid Y$) to treat truncation in $Y$. Although the inverse regression is insensitive to the truncation, it estimates a completely different function of the underlying population and, under realistic situations with small data sets and correlation coefficients less than 1, is numerically the least stable of all OLS lines (BF).

Computer software implementing truncated regression for bivariate normal distributions can be purchased; the econometrics programs SHAZAM (White et al. 1988) and especially LIMDEP (Greene 1989) provide a wide range of truncated and sample selection regression models.

## 6. CENSORED REGRESSION MODELS

Censored data sets are those where the exact values of certain points are known only to be below (or above) some limiting value(s). They occur frequently in astronomy when a survey of a new property is made of a preselected sample of objects, and some objects are not detected. Analysis of censored data is often more effective than analysis of truncated data: with censoring, something is known about the location of all points in the sample, while with truncation nothing at all is known about the number or location of points beyond the truncation limit. Linear regression with censoring in one or both variables has been developed primarily within the field of "survival analysis" for biometrics (e.g., Buckley & James 1979; Miller & Halpern 1982) and in econometrics, where it is known as the Tobit regression model (see references in § 5). In astronomical context, censored regression has been discussed by Schmitt (1985) and Isobe, Feigelson, & Nelson (1986). The reader is referred to these papers and the reviews of Feigelson (1990, 1992) for further information.

ASURV, a FORTRAN software package implementing three censored regression models—the EM algorithm assuming Gaussian residuals, the Buckley-James line, and Schmitt's binned regression—is being distributed within the astronomical community (Isobe & Feigelson 1990). A revised and enlarged version (REV 1) is now available (LaValley, Isobe, & Feigelson 1992) and can be obtained from the authors (Internet: code@stat.psu.edu).

## 7. EXAMPLES OF ASTRONOMICAL APPLICATIONS

### 7.1. *Unweighted Linear Regression:* log $D_n$–log $\sigma$ *Relation in Elliptical Galaxies*

The revised Faber-Jackson relation involving log $D_n$ and log $\sigma$, where $D_n$ is a linear diameter based on optical surface brightness and $\sigma$ is the stellar velocity dispersion, provides accurate distances to elliptical galaxies. These are a crucial link in the study of large-scale galaxy streaming and the Great Attractor. Table 1 and Figure 7 show a data set of 18 Virgo Cluster elliptical galaxies from Dressler et al. (1987). They report a "median" fitted line $\log \sigma = 0.750 \log D_n + 0.934$, and its inverted relation $\log D_n = 1.333 \log \sigma - 1.237$, where the "median" line is what we call the OLS bisector (Lynden-Bell et al. 1988, Appendix D). No error analysis is provided, but application of the standard asymptotic normal OLS($Y \mid X$) formulae (e.g., Bevington 1969) gives intercept and slope uncertainties of $\pm 0.078$ and $\pm 0.177$, respectively.

TABLE 1

log $D_n$ VERSUS log $\sigma$ FOR VIRGO ELLIPTICAL GALAXIES[a]

| log $D_n$ | log $\sigma$ | log $D_n$ | log $\sigma$ |
|---|---|---|---|
| 1.868 | 2.412 | 1.678 | 2.170 |
| 2.028 | 2.480 | 2.068 | 2.528 |
| 1.478 | 2.059 | 1.488 | 2.021 |
| 1.978 | 2.355 | 1.928 | 2.391 |
| 1.448 | 2.009 | 1.668 | 2.185 |
| 1.378 | 1.949 | 1.908 | 2.338 |
| 1.418 | 2.079 | 1.858 | 2.303 |
| 2.118 | 2.474 | 2.078 | 2.514 |
| 1.858 | 2.268 | 1.688 | 2.262 |

[a] Data obtained from Table 2 of Dressler et al. 1987 with outliers V1 and V13 omitted.

Table 2 shows the application of our six lines. The sample is small ($N = 18$), but the linear correlation coefficient is high ($\rho = 0.97$). The former implies that the analytical error analyses are likely to be too small, while the latter implies that the calculated slopes should be quite close together. Since the goal here is to evaluate galaxy distances, we recommended use of the standard OLS($Y \mid X$) line where the distance-dependent quantity (log $D_n$) is the $Y$ variable. The result is $\log D_n = (1.30 \pm 0.07) \log \sigma - (1.17 \pm 0.16)$, where the uncertainties are the largest of the asymptotic, bootstrap, and jacknife computations. While the use of the OLS bisector by Dressler et al. (1987) for this regression is not mathematically incorrect, it seems ill-advised for two reasons: the goal of the experiments is to predict distance-related values, so minimizing residuals in $D_n$ is warranted; if the bisector is used for this initial regression, it must be self-consistently used in all later stages of the complex analysis (cf. Lynden-Bell et al. 1988) leading to galaxy streaming and other results. In any case, the regression coefficient uncertainties should be propagated through all later calculations.

### 7.2. *The Calibration Problem: Distance between Coma and Virgo Clusters*

Much effort has been directed to establishing the relative distances of clusters of galaxies using empirical correlations

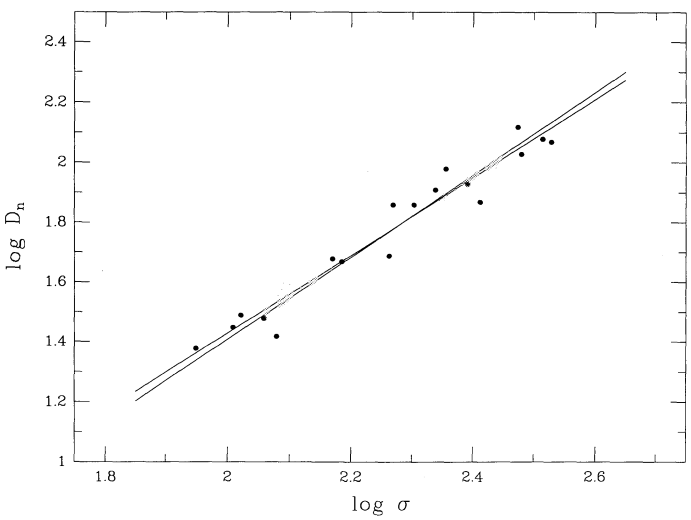


FIG. 7.—Plot of central velocity dispersion and diameter measure for a sample of Virgo Cluster elliptical galaxies (Dressler et al. 1987), with the two OLS lines shown.

TABLE 2

REGRESSIONS FOR log $D_n$ VERSUS log $\sigma$

| METHOD (1) | ASYMPTOTIC FORMULAE Intercept (2) | ASYMPTOTIC FORMULAE Slope (3) | BOOTSTRAP SLOPE (4) | JACKKNIFE SLOPE (5) |
|---|---|---|---|---|
| OLS($Y \mid X$) | $-1.171 \pm 0.136$ | $1.300 \pm 0.060$ | $1.311 \pm 0.071$ | $1.300 \pm 0.067$ |
| OLS($X \mid Y$) | $-1.340 \pm 0.170$ | $1.374 \pm 0.075$ | $1.389 \pm 0.087$ | $1.374 \pm 0.085$ |
| OLS bisector | $-1.254 \pm 0.150$ | $1.336 \pm 0.066$ | $1.349 \pm 0.077$ | $1.336 \pm 0.075$ |
| Orthogonal | $-1.279 \pm 0.159$ | $1.347 \pm 0.070$ | $1.361 \pm 0.082$ | $1.347 \pm 0.079$ |
| Reduced major axis | $-1.255 \pm 0.150$ | $1.336 \pm 0.066$ | $1.349 \pm 0.077$ | $1.337 \pm 0.075$ |
| Mean OLS | $-1.256 \pm 0.150$ | $1.337 \pm 0.067$ | $1.350 \pm 0.077$ | $1.337 \pm 0.075$ |

NOTE.—Asymptotic formula results are based on the equations given in Table 1 in IFAB and BF. All results are from the program SLOPES.

Col. (2).—Asymptotic formula intercept and standard deviation.
Col. (3).—Asymptotic formula slope and standard deviation.
Col. (4).—Mean and standard deviation of 200 bootstrap resamplings of the data.
Col. (5).—Mean and standard deviation of $N$ jackknife resamplings of the data.

between distance-independent (e.g., global rotational velocities or stellar velocity dispersions) and distance-dependent (e.g., absolute luminosities or linear diameters) variables. One recent effort by Lucey, Bower, & Ellis (1991, hereafter LBE) discusses the distance offset between the Coma and Virgo clusters using a revised Faber-Jackson relation for elliptical galaxies similar to that used by Dressler et al. (1987), presented above. LBE show that Coma elliptical galaxies occupy a "fundamental plane" in velocity dispersion–diameter–surface brightness space, with no scatter beyond that attributable to measurement errors. To obtain the fit, they "minimized the log $\sigma$, log $A_e$ and $\mu_e$ residuals independently and averaged the coefficients." This resembles our OLS mean procedure, but treats log $\sigma$ as the independent variable although the distance-dependent diameter is the variable being estimated.

Adopting their definition $D'_n$ of galaxy diameter that includes a small surface brightness term, their Coma relation reduces to the bivariate relation log $D'_n \sim 1.215$ log $\sigma$. LBE then apply this calibration relation to 16 Virgo elliptical galaxies and derive a Coma-Virgo distance offset of $\Delta \log D'_n = 0.765 \pm 0.021$. Reversing the sequence and applying a Virgo-based calibration with a slope of 1.200 (obtained from a different regression analysis of several clusters) to the Coma galaxies gives a somewhat different offset of $0.729 \pm 0.025$. These offsets and uncertainties are based on an application of the Wilcoxon rank sum test (Dressler 1984). LBE conclude that, because the use of either cluster as calibrator is equally valid and the results differ, the cluster distance offset cannot be determined to better than 9% accuracy.

We can address this problem using the calibration methods described in § 3 above. The data are given in the "observed" columns of Table 3 and are plotted in Figure 8. Our regression solutions (Table 4) show that the slopes $\beta = 1.200$ and $\beta = 1.215$ used by LBE lie between the slopes of the two OLS extrema, but are not consistent solutions of any of the six regression procedures. We take a single regression procedure, the OLS(log $D'_n$|log $\sigma$) line, chosen because it minimizes the residuals of the distance-dependent variable log $D'_n$ and should thus give the least scatter in the resulting distance offset.

First, we apply the offset formulae (A4)–(A8) in the Appendix, which weights all points identically, shifting the log $\sigma$ axis to the mean of the calibration sample temporarily for the calculation. Taking the Coma sample as calibrator with $\beta \pm \sigma_\beta = 1.162 \pm 0.111$ (using the jackknife rather than asymptotic errors because the sample is small), the intercepts are $\alpha(\text{Coma}) = 1.034 \pm 0.008$ and $\alpha(\text{Virgo}) = 1.798 \pm 0.021$. The intercept offset is found to be $0.764 \pm 0.021$. Nearly identical values are obtained when the calibrator and new samples are switched: $\beta \pm \sigma_\beta = 1.144 \pm 0.118$ from Virgo, giving offset $0.764 \pm 0.023$. No incompatibility in the cross-calibrations is

TABLE 3

log $D'_n$ VERSUS log $\sigma$ FOR COMA AND VIRGO ELLIPTICAL GALAXIES[a]

| COMA Observed log $\sigma$ | COMA Observed log $D'_n$ | COMA Predicted log $D'_n$ | VIRGO Observed log $\sigma$ | VIRGO Observed log $D'_n$ | VIRGO Predicted log $D'_n$ |
|---|---|---|---|---|---|
| 2.391 | 1.222 | $1.946 \pm 0.031$ | 2.412 | 1.814 | $1.208 \pm 0.018$ |
| 2.300 | 1.049 | $1.842 \pm 0.026$ | 2.480 | 2.054 | $1.288 \pm 0.025$ |
| 2.175 | 0.932 | $1.699 \pm 0.029$ | 2.059 | 1.497 | $0.798 \pm 0.026$ |
| 2.287 | 1.106 | $1.827 \pm 0.026$ | 2.355 | 1.889 | $1.142 \pm 0.013$ |
| 2.223 | 1.012 | $1.754 \pm 0.027$ | 2.009 | 1.421 | $0.740 \pm 0.031$ |
| 2.248 | 0.997 | $1.782 \pm 0.026$ | 2.079 | 1.629 | $0.821 \pm 0.023$ |
| 2.221 | 0.977 | $1.751 \pm 0.027$ | 2.474 | 2.102 | $1.281 \pm 0.024$ |
| 2.325 | 1.155 | $1.870 \pm 0.027$ | 2.268 | 1.943 | $1.041 \pm 0.009$ |
| 2.345 | 1.128 | $1.893 \pm 0.028$ | 2.170 | 1.802 | $0.927 \pm 0.015$ |
| 2.261 | 0.980 | $1.797 \pm 0.026$ | 2.528 | 2.004 | $1.343 \pm 0.030$ |
| 2.581 | 1.357 | $2.163 \pm 0.050$ | 2.021 | 1.484 | $0.754 \pm 0.029$ |
| 2.023 | 0.808 | $1.525 \pm 0.042$ | 2.391 | 2.027 | $1.184 \pm 0.016$ |
| 2.197 | 0.978 | $1.724 \pm 0.028$ | 2.185 | 1.691 | $0.945 \pm 0.014$ |
| 2.130 | 0.841 | $1.647 \pm 0.032$ | 2.338 | 1.922 | $1.122 \pm 0.012$ |
| 2.274 | 1.084 | $1.812 \pm 0.026$ | 2.514 | 2.059 | $1.327 \pm 0.028$ |
| 2.323 | 1.129 | $1.868 \pm 0.027$ | 2.262 | 1.838 | $1.034 \pm 0.009$ |
| 2.199 | 0.977 | $1.726 \pm 0.027$ | | | |
| 2.252 | 1.048 | $1.787 \pm 0.026$ | | | |
| 0.763 | 2.079 | $1.589 \pm 0.036$ | | | |
| 1.156 | 2.373 | $1.925 \pm 0.030$ | | | |
| 0.876 | 2.167 | $1.690 \pm 0.029$ | | | |
| 1.026 | 2.270 | $1.807 \pm 0.026$ | | | |
| 1.181 | 2.380 | $1.933 \pm 0.030$ | | | |

[a] Coma values are derived from Table 1 of Lucey, Bower, & Ellis 1991, with $D'_n = D_n - 0.117[\mu_e(b) - 21.0]$ and three galaxies having $\mu_e > 22.0$ excluded. Virgo values are from Table 2 of Dressler et al. 1987, with $D'_n = D_n - 0.117[\Sigma_e(b) - 21.0]$, two galaxies excluded as outliers and three galaxies excluded for $\Sigma_e > 22.0$. The "predicted" Coma values are derived by applying the Virgo OLS(log $D'_n$| log $\sigma$) calibration line and the formulae in § A2 of the Appendix, and the "predicted" Virgo values are derived similarly from the Coma OLS(log $D'_n$|log $\sigma$) calibration line. The uncertainties of the predicted values have been increased by a factor of 1.35 over the analytical values, to account for a similar factor of the jackknife uncertainties over the asymptotic uncertainties in the calibration regressions (see Table 4).

TABLE 4

REGRESSIONS FOR COMA AND VIRGO log $D_n'$ VERSUS log $\sigma$[a]

| METHOD (1) | ASYMPTOTIC FORMULAE Intercept (2) | ASYMPTOTIC FORMULAE Slope (3) | BOOTSTRAP SLOPE (4) | JACKKNIFE SLOPE (5) |
|---|---|---|---|---|
| 23 Coma Ellipticals | | | | |
| OLS($Y \mid X$) | $-1.595 \pm 0.186$ | $1.162 \pm 0.082$ | $1.186 \pm 0.094$ | $1.164 \pm 0.111$ |
| OLS($X \mid Y$) | $-1.765 \pm 0.216$ | $1.238 \pm 0.096$ | $1.261 \pm 0.104$ | $1.239 \pm 0.128$ |
| OLS bisector | $-1.678 \pm 0.200$ | $1.199 \pm 0.088$ | $1.223 \pm 0.099$ | $1.201 \pm 0.119$ |
| Orthogonal | $-1.694 \pm 0.209$ | $1.206 \pm 0.092$ | $1.231 \pm 0.102$ | $1.208 \pm 0.124$ |
| Reduced major axis | $-1.679 \pm 0.200$ | $1.199 \pm 0.088$ | $1.223 \pm 0.099$ | $1.201 \pm 0.119$ |
| OLS mean | $-1.680 \pm 0.200$ | $1.200 \pm 0.088$ | $1.224 \pm 0.099$ | $1.201 \pm 0.119$ |
| 16 Virgo Ellipticals | | | | |
| OLS($Y \mid X$) | $-0.790 \pm 0.230$ | $1.144 \pm 0.101$ | $1.143 \pm 0.127$ | $1.114 \pm 0.118$ |
| OLS($X \mid Y$) | $-1.183 \pm 0.180$ | $1.316 \pm 0.082$ | $1.322 \pm 0.132$ | $1.316 \pm 0.093$ |
| OLS bisector | $-0.978 \pm 0.190$ | $1.227 \pm 0.085$ | $1.227 \pm 0.107$ | $1.226 \pm 0.099$ |
| Orthogonal | $-1.021 \pm 0.198$ | $1.245 \pm 0.089$ | $1.246 \pm 0.121$ | $1.245 \pm 0.104$ |
| Reduced major axis | $-0.979 \pm 0.190$ | $1.227 \pm 0.085$ | $1.228 \pm 0.108$ | $1.227 \pm 0.099$ |
| OLS mean | $-0.986 \pm 0.188$ | $1.230 \pm 0.084$ | $1.233 \pm 0.110$ | $1.230 \pm 0.098$ |

[a] See footnote to Table 2.

present. We make two parenthetical comments. First, the intercept offset is virtually unchanged ($<0.002$ difference) when the other five regression lines are used, provided that a single method is used throughout the calculation. Second, the offset uncertainty was only slightly affected by the error in the calibration slope in this example, because the calibration and new samples happened to have nearly identical mean abscissa values (i.e., the last term in eq. [A4] happened to be small). In other cases this term can dominate the calculated offset uncertainty.

Next, we adopt the procedure discussed in § 3.2, based on curved confidence bands, which recognizes that the calibration regression's accuracy is reduced at the ends of the relation. The columns of Table 3 labeled "predicted" diameters and uncertainties, calculated individually for each Virgo (Coma) galaxy using the OLS(log $D_n'$|log $\sigma$) Coma (Virgo) calibration line described in § A2 of the Appendix. The difference between the predicted and observed diameters for each galaxy is a measure of the distance offset between the Coma and Virgo clusters. Table 5 gives both unweighted and weighted means of these differences. Weights are the quadratic sums of the calibration uncertainties for the predicted values and the measurement errors $\Delta(\log D_n') = 0.011$ estimated by LBE from replicated observations. Note that the uncertainties of the means are $N^{-1/2}$ times the standard deviation of the sample.

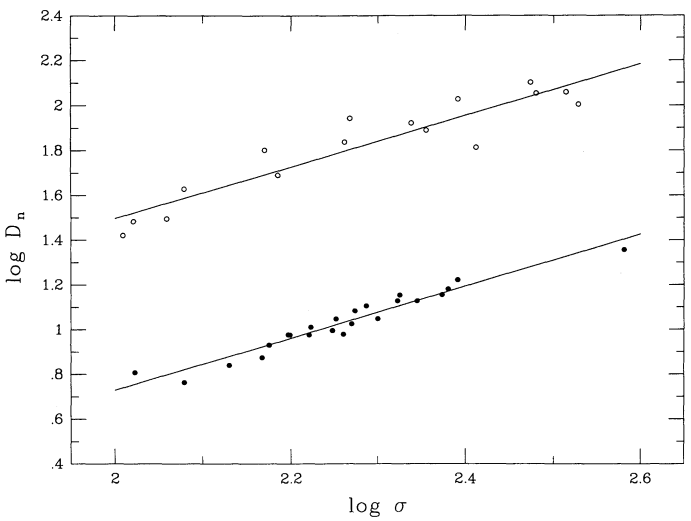


FIG. 8.—Plot of the central velocity dispersion and revised diameter measure for Virgo (*open circles*) and Coma (*filled circles*) elliptical galaxies (Lucey et al. 1991). OLS(log $D_n'$|log $\sigma$) lines are shown. See text for calibration procedures

The unweighted mean (Virgo − Coma) offsets are virtually identical to those derived by the method above, though with smaller uncertainties. Similar calculations using the other least-squares regression lines give negligibly different simple mean offsets (0.760–0.764). The weighted offset with Coma as the calibrator is a poor solution because the predicted diameter uncertainties (Table 4) are much smaller than the sample standard deviation (Table 5), but the weighted solution with Virgo as the calibrator is satisfactory. We adopt the most precise result, $(\Delta D_n') = 0.763 \pm 0.006$. All less precise solutions, independent of whether Virgo or Coma galaxies are considered the calibrator sample, are completely consistent with this one.

We conclude that the data of LBE permit the distance offset between the Coma and Virgo clusters to be determined to 1% accuracy: $\Delta \log D_n' = 0.764 \pm 0.021$ using the intercept offset discussed in § 3.1, and $0.763 \pm 0.006$ using the generalized Working-Hotelling confidence bands presented in § 3.2. The latter method is more precise because the different weights of the points at different locations along the regression line are taken into account.

TABLE 5

RESULTS FOR VIRGO − COMA DISTANCE OFFSET

| Calibrator Sample | Simple Mean Offset ($\Delta D_n'$) | Weighted Mean Offset ($\Delta D_n'$) | Standard Deviation |
|---|---|---|---|
| Coma | $0.7638 \pm 0.0206$ | $0.7868 \pm 0.0052$ | 0.0797 |
| Virgo | $0.7642 \pm 0.0074$ | $0.7626 \pm 0.0064$ | 0.0345 |

## 8. DISCUSSION AND CONCLUSIONS

While linear regression with two variables appears at first to be a simple statistical problem, we hope to have communicated here and in our earlier papers (IFAB; BF) that the issues are surprisingly complex. Astronomers are not the only group of scholars to confront them: vociferous (and sometimes inconclusive) debates on such problems have occurred among biometricians, econometricians, and statisticians over the past century. Following are our conclusions; some are stated with considerable confidence, while others only raise issues to be considered by the wider astronomical community.

1. Astronomers should select their regression method on a case-by-case basis, considering any available knowledge beyond the $(x_i, y_i)$ values (such as measurement errors, truncation, or censoring limits) and the precise scientific question. Careful enunciation of the question is important, for example, when studying the expansion of the universe from a sample of galaxies. One might ask: Are we seeking the best estimate of $H_0$ defined by Hubble's law $v = H_0 d + v_0$, where all of the scatter arises in the velocities and none in the distances? Are we seeking the best estimate of some "structural" relationship between velocity and distance that makes no judgment on whether velocity depends on distance or vice versa? Are we seeking the best estimate of the age of the universe, which is proportional to $1/H_0$? In the first case (assuming equally weighted data points) one might calculate OLS($v \mid d$), in the second case one might calculate the OLS bisector, and in the third case one might reverse the variables and calculate OLS($d \mid v$).

Much of the proliferation of linear regression methods within the cosmic distance scale literature is due to imprecise definition, or to the lack of consensus on the precise definition, of the scientific question being addressed. This difficulty of transforming a scientific issue to precise mathematical questions is not unique to astronomy and can only be resolved by agreement among researchers.[4] Once the scientific question and assumptions are precisely formulated, then many statistical treatments are clearly excluded.

For the simple linear regression problem described in § 2 (unweighted, uncensored, untruncated but without clear causal relations between $X$ and $Y$), the following statistical characteristics of the problem might be considered. If one variable is known to have negligible variance, either intrinsic to the objects or due to the measurement process, and all of the scatter is in the other variable, then the latter generally should be the dependent variable and the standard OLS($Y \mid X$) regression should be performed. If the scientific question being addressed can be unambiguously phrased as "How does property $A$ depend on property $B$?" or "how can we predict values of $A$ for new objects given measurement of $B$?" then property $A$ should be the dependent variable and OLS($A \mid B$) should be performed. If these conditions do not hold, or if the question can be definitely phrased as "What is the intrinsic relation between properties $A$ and $B$ in these objects, without treating one variable differently than the other?" then we advocate use of a regression method that treats the variables symmetrically.

2. We reiterate that the six unweighted regression lines are intrinsically different functionalities of the data set under study and are *not interchangeable*. Astronomers studying the cosmic distance scale should be no more readily to switch linear fits than they are to interchange linear and nonlinear fits.

3. When an unweighted method treating the variables symmetrically is desired, the OLS bisector is to be preferred. This is a reiteration of the conclusion reached in IFAB and BF. It is based not on a mathematical demonstration that this is a "better fit" than the other three symmetrical regressions, since they are all mathematically valid functions of the data. Rather, the recommendation is based on the findings that the OLS bisector (*a*) has a smaller variance about its theoretical fit than the other lines and (*b*) does not share the mathematical limitations of some other symmetrical fits.

4. The inverse OLS($X \mid Y$) line should be avoided, since it has the greatest variance and the least stability in achieving its theoretical regression coefficients. If the $Y$ variable is truncated so that the standard OLS($Y \mid X$) is biased, then either the variables should be switched so that a standard OLS($Y \mid X$) can be used, or a regression designed to compensate for the truncation bias (see § 6) should be used.

5. The standard error analysis for the OLS($Y \mid X$) regression coefficients appearing in elementary statistics texts is mathematically incorrect for any other regression line and can be inaccurate even for OLS($Y \mid X$) under many conditions. Asymptotic analytic expressions for estimates of variances of slopes and intercepts for the six regression lines given in IFAB and BF are valid even when the distributions are not normal and the $X$-values have scatter. For small samples, numerical resampling and simulations should be performed, as described in § 2 above and implemented in the FORTRAN program SLOPES.

6. Slope estimates based on the average of the ratios $y_i/x_i$ should not be performed, except under very specific circumstances.

7. If the uncertainties in one or both variables are known from detailed knowledge of the measurement process, and if these measurement errors dominate the scatter around a regression line, then (doubly) weighted functional regression procedures should be used. Simple regression formulae exist for homoscedastic measurement error models, but iterative methods are needed for heteroscedastic problems. Several computer codes are available. If measurement errors are present, but contribute insignificantly to the scatter, then one of the six lines described in § 2 should be used. Models that incorporate both measurement errors and intrinsic scatter are complex and not yet fully developed.

8. When a series of regressions is performed, as in cosmic distance calculations based on calibration regressions applied to new samples, the errors in both slopes and intercepts of the calibration line should be propagated. One approach is to view the two samples in their entireties, and estimate the intercept offsets (§ 3.1 and § A1 of the Appendix). But it is preferable to treat the second sample as a collection of individual measurements of the independent variable. The calibration regression is then applied to predict new dependent variable values, with the error analysis varying from point to point (§ 3.2 and § A2 of the Appendix). A code implementing this approach (CALIB) is

[4] A recent controversy in economics is quite instructive in this regard (see Conway & Roberts 1983, and the discussions it engendered in the 1984 April issue of the Journal of Business & Economic Statistics). To determine whether an employer unfairly discriminates against women, some economists advocate a regression of salary versus gender, while others advocate a regression of gender versus salary. The results can be very different, and they have led to lawsuits that have reached the highest levels of federal courts. It appears that the choice requires a consensus on the precise meaning of the concept "fairness in employment," which does not currently exist.

available from the authors. Even for the standard OLS($Y \mid X$) line, our solution should be preferred over the standard Working-Hotelling confidence bands in most astronomical applications, because our solution permits the $X$-values to be random variables rather than fixed quantities.

We thank R. Ciardullo (Penn State) for helpful discussions concerning the manuscript, F. Bookstein (Michigan) for a stimulating conversation on structural models, F. Murtagh (ST–SCF) for encouragement and software, and T. Isobe (MIT) for contributions to early stages of this project. The anonymous referee made helpful comments and located an error in the original manuscript. This work was founded by NASA grants NAGW-1917, NAGW-2120, and by NSF grant DMS-9007717.

## APPENDIX

## MATHEMATICAL DERIVATIONS OF CALIBRATION ERRORS

### A1. INTERCEPT OF A CALIBRATOR LINE APPLIED TO A NEW SAMPLE

Suppose that $W$ is a random variable with expected value $E(W) = \mu$, and $Z$ is any covariate. The variance of $W$ can be expressed as a function of conditional mean and variance given $Z$:

$$\begin{aligned} \text{Var}\,(W) = E(W-\mu)^2 &= E[W - E(W \mid Z)]^2 + E[E(W \mid Z) - \mu]^2 \\ &= E[\text{Var}\,(W \mid Z)] + \text{Var}\,[W(W \mid Z)]\,. \end{aligned} \tag{A1}$$

We can now apply this principle to obtain the variance of the estimate of intercept $\alpha$ for a new bivariate sample $(x_i, y_i)$, $i = 1, \ldots, N$, given the slope $\beta_{\rm cal}$ and slope uncertainty $\sigma(\beta_{\rm cal})$ derived from a regression of an independent calibrator sample. Let

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad S_{xy} = \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\,; \qquad S_{xx} = \sum_{i=1}^{N}(x_i - \bar{x})^2 \qquad \text{and} \qquad S_{yy} = \sum_{i=1}^{N}(y_i - \bar{y})^2\,. \tag{A2}$$

Then the $y$ intercept is estimated by

$$\hat{\alpha} = \bar{y} - \beta_{\rm cal}\bar{x}\,, \tag{A3}$$

and its variance is (after some algebra not shown) given by

$$\text{Var}\,(\hat{\alpha}) = \sigma^2(\beta_{\rm cal})E(\bar{x}^2) + \frac{1}{N}\,\text{Var}\,[Y - E(\beta_{\rm cal})X]\,.$$

By substituting sample quantities for population quantities, we derive the following estimate of the variance of the intercept:

$$\widehat{\text{Var}}\,(\hat{\alpha}) = [S_{yy} + \sigma^2(\beta_{\rm cal})S_{xx} + \beta_{\rm cal}^2 S_{xx} - 2S_{xy}\beta_{\rm cal}]N^{-2} + \sigma^2(\beta_{\rm cal})\bar{x}^2\,. \tag{A4}$$

Since the entire regression calculation applies for arbitrary translations along the two axes, the abscissa zero point can be set equal to the mean abscissa value of the calibration sample, $\bar{x}_{\rm cal}$. The last term in equation (A4) will then be negligible if the calibration and new samples have nearly identical distributions of $x$-values.

Astronomers are often interested in the offset between the intercepts of the calibration and new samples. This offset is given by

$$\hat{\alpha} - \alpha_{\rm cal} = \bar{y} - \bar{y}_{\rm cal} - \beta_{\rm cal}(\bar{x} - \bar{x}_{\rm cal})\,. \tag{A5}$$

The variance estimate of this offset is then

$$\widehat{\text{Var}}\,(\hat{\alpha} - \alpha_{\rm cal}) = \widehat{\text{Var}}\,(\hat{\alpha}) + \widehat{\text{Var}}\,(\alpha_{\rm cal}) - 2\,\widehat{\text{Cov}}\,(\hat{\alpha}, \alpha_{\rm cal})\,, \tag{A6}$$

where the first term is from equation (A4), the second term is the intercept variance of the selected OLS line given in IFAB and BF (see § 2 above), and the third term can be estimated by

$$\widehat{\text{Cov}}\,(\hat{\alpha}, \alpha_{\rm cal}) = -\bar{x}\sum_{i=1}^{N_{\rm cal}}\{(y_i^c - \beta_{\rm cal}\,x_i^c)[\hat{a}^c(y_i^c - \beta_{1,\rm cal}\,x_i^c)x_i^c + \hat{b}^c(y_i^c - \beta_{2,\rm cal}\,x_i^c)y_i^c]\} + \bar{x}\bar{x}_{\rm cal}\,\sigma^2(\beta_{\rm cal})\,. \tag{A7}$$

Here $N_{\rm cal}$ is the calibration sample size, $x_i^c = x_{i,\rm cal} - \bar{x}_{\rm cal}$, $y_i^c = y_{i,\rm cal} - \bar{y}_{\rm cal}$, $\beta_{1,\rm cal}$ and $\beta_{2,\rm cal}$ are the OLS($Y \mid X$) and OLS($X \mid Y$) estimated slopes for the calibrator sample, $a^c$ and $b^c$ are given in Table 6 for the selected OLS line with $\hat{\beta}_j$ values calculated for the calibration sample, and the constants $\hat{\psi}$ and $\hat{\omega}$ are defined to be

$$\hat{\psi} = \left[N_{\rm cal}\sum_{i=1}^{N_{\rm cal}}(x_i^c)^2\right]^{-1}, \qquad \hat{\omega} = \left(N_{\rm cal}\sum_{i=1}^{N_{\rm cal}} x_i^c\,y_i^c\right)^{-1}, \tag{A8}$$

TABLE 6

PARAMETERS $\hat{a}_j$ AND $\hat{b}_j$ FOR CALIBRATION VARIANCES

| $j$ | $\hat{a}_j$ | $\hat{b}_j$ |
|---|---|---|
| 1 | $\hat{\psi}$ | 0 |
| 2 | 0 | $\hat{\omega}$ |
| 3 | $\hat{\psi}\hat{\beta}_3[(1+\hat{\beta}_2^2)/(1+\hat{\beta}_1^2)]^{1/2}(\hat{\beta}_1+\hat{\beta}_2)^{-1}$ | $\hat{\omega}\hat{\beta}_3[(1+\hat{\beta}_1^2)/(1+\hat{\beta}_2^2)]^{1/2}(\hat{\beta}_1+\hat{\beta}_2)^{-1}$ |
| 4 | $\hat{\psi}(\hat{\beta}_4/\lvert\hat{\beta}_1\rvert)[4\hat{\beta}_1^2+(\hat{\beta}_1\hat{\beta}_2-1)^2]^{-1/2}$ | $\hat{\omega}\hat{\beta}_4\lvert\hat{\beta}_1\rvert[4\hat{\beta}_1^2+(\hat{\beta}_1\hat{\beta}_2-1)^2]^{-1/2}$ |
| 5 | $\hat{\psi}(\hat{\beta}_2/4\hat{\beta}_1)^{1/2}$ | $\hat{\omega}(\hat{\beta}_1/4\hat{\beta}_2)^{1/2}$ |
| 6 | $\frac{1}{2}\hat{\psi}$ | $\frac{1}{2}\hat{\omega}$ |

## A2. GENERALIZED CONFIDENCE STRIPS

The standard calibration problem based on Working & Hotelling (1929; see also other references in § 3.2) calculates the 90% (or other) confidence level curves around a linear regression, representing the uncertainty of the estimate of ordinate values given a new abscissa value. The derivation of the standard result, however, assume that the regression is OLS$(Y \mid X)$ and that the $X$-values are known precisely. This is a reasonable assumption in, say, analytical chemistry, where the calibration experiment can be precisely controlled. But it clearly is not a good assumption in astronomy, where the calibrating stars or galaxies are selected from the large population simply by virtue of their proximity to the observer, and will have the same random properties as any other star or galaxy. We therefore derive here a generalization of Working-Hotelling confidence strips that is applicable when both variables are random, and when any of the six least-squares regressions are performed.

Let $(x_i, y_i), i = 1, \ldots, N$ be independent observations from a common population with mean $(\mu_x, \mu_y)$ and covariance matrix

$$S = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} . \tag{A9}$$

Consider this to be the calibration sample. Let $x_0$ be a new fixed point. For each of the six unweighted least-squares linear regression procedures described in IFAB, $j = 1, \ldots, 6$, we seek the values of the dependent variable corresponding to $x_0$,

$$\begin{aligned} \hat{y}_{0j} &= \hat{\alpha}_j + x_0\hat{\beta}_j \\ &= \bar{y} - \beta_j\bar{x} + (x_0 - \mu_x)\hat{\beta}_j + (\beta_j - \hat{\beta}_j)(\bar{x} - \mu_x) + \mu_x\beta_j . \end{aligned} \tag{A10}$$

The values of $\hat{\alpha}_j$ and $\hat{\beta}_j$ for $j = 1, \ldots, 5$ are given in Table 1 and equation (8) of IFAB. The sixth value, $\hat{\beta}_6 = \frac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2)$ and the corresponding intercept, is given in BF. Following Lemma A1 and Part II of Appendix A of IFAB, after some algebra the asymptotic variance $\sigma_{0j}^2$ of $\hat{y}_{0j}$ can be expressed as

$$\sigma_{0j}^2 = N^{-1}\ \mathrm{Var}\ \{Y - \beta_j X + a_j(X - \mu_x)[Y - \mu_y - \beta_1(X - \mu_x)] + b_j(Y - \mu_y)[Y - \mu_y - \beta_2(X - \mu_x)]\} . \tag{A11}$$

Substituting sample quantities for population quantities, $\sigma_{0j}^2$ can be estimated by

$$\hat{\sigma}_j^2 = \frac{1}{N^2}\sum_{i=1}^{N}\{y_i - \bar{y} - \hat{\beta}_j(x_i - \bar{x}) + \hat{a}_j(x_i - \bar{x})[y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})] + \hat{b}_j(y_i - \bar{y})[y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x})]\}^2 . \tag{A12}$$

The values $\hat{a}_j$ and $\hat{b}_j$ are given in Table 6 (the corresponding population values of $a_j$ and $b_j$ can be similarly expressed), and the constant $\hat{\psi}$ and $\hat{\omega}$ are defined here to be

$$\hat{\psi} = N(x_0 - \bar{x})/S_{xx} , \qquad \hat{\omega} = N(x_0 - \bar{x})\hat{\beta}_2/S_{yy} . \tag{A13}$$

This solution can also be readily derived from equations (9)–(21) in IFAB.

## REFERENCES


Aaronson, M., Bothun, G., Mould, J., Huchra, J., Schommer, R. A., & Cornell, M. E. 1986, ApJ, 302, 536
Akritas, M. G., Saleh, A. K., & Sen, P. K. 1985, in Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences, ed. P. K. Sen (New York: Elsevier)
Amemiya, T. 1985, Advanced Econometrics (Cambridge: Harvard Univ. Press)
Anderson, R. L. 1987, Practical Statistics for Analytical Chemists (New York: Van Nostrand Reinhold)
Anderson, T. W. 1984, Ann. Statist., 12, 1
Anderson, T. W., & Sawa, T. 1982, J. R. Stat. Soc. B, 44, 52
Antille, A., & Milasevic, P. 1988, Prob. Theor. Rel. Fields, 78, 63
Babu, G. J. 1992, in Statistics '91 Canada, in press
Babu, G. J., & Feigelson, E. D. 1992, Commun. Statist. Comput. Simul., 22, 533 (BF)
Bevington, P. R. 1969, Data Reduction and Error Analysis for the Physical Sciences (New York: McGraw-Hill)
Bhattacharya, P. K. 1983, Ann. Statist., 11, 697
Bhattacharya, P. K., Chernoff, H., & Yang, S. S. 1983, Ann. Statist., 11, 505
Bickel, P. J., & Ritov, J. 1991, Ann. Statist., 19, 797
Boggs, P. T., Byrd, R. H., & Schnabel, R. B. 1987, SIAM J. Sci. Stat. Comput., 8, 1052
Boggs, P. T., Donaldson, J. R., Byrd, R. H., & Schnabel, R. B. 1990, ACM Trans. Math. Software, 15, 348
Boswell, M. 1990, TULSIM, Rev. 3.2 (computer program) (available from M. T. Boswell, Department of Statistics, Penn State University, University Park, PA 16802)
Branham, R. L., Jr. 1982, AJ, 87, 928
Buckley, J., & James, I. 1979, Biometrika, 66, 429
Cameron, J. M. 1982, in Encyclopedia of Statistical Sciences, 1, 346
Conway, D. A., & Roberts, H. V., 1983, J. Bus. Econ. Statist., 1, 75
Deeming, T. J. 1968, Vistas Astron., 10, 125
Deming, W. E. 1943, Statistical Adjustment of Data (New York: Wiley)
Dressler, A. 1984, ApJ, 281, 512
Dressler, A., Lynden-Bell, D., Burstein, D., Davies, R. L., Faber, S. M., Terlevich, R. J., & Wegner, G. 1987, ApJ, 313, 42
Efron, B., & Tibshirani, R. 1986, Stat. Science 1, 54

Fasano, G., & Vio, R. 1988, Bull. Inf. Cent. Données Stellaires, 35, 191
Feigelson, E. D. 1990, in Errors, Bias, and Uncertainties in Astronomy, ed. C. Jaschek & F. Murtagh (Cambridge: Cambridge Univ. Press), 213
———. 1992, in Statistical Challenges in Modern Astronomy, ed. E. D. Feigelson & G. J. Babu (New York: Springer-Verlag)
Fouqué, P., Bottinelli, L., Gougenheim, L., & Paturel, G. 1990, ApJ, 349, 1
Fuller, W. A. 1987, Measurement Error Models (New York: Wiley)
———. 1988, EV CARP (computer program) (available from Statistical Laboratory, Iowa State University, Ames, IA 50011)
Greene, W. H. 1989, LIMDEP (computer program) (available from Econometric Software, Inc., 43 Maple Avenue, Bellport, NY 11713)
———. 1991, Econometric Analysis (New York: Macmillan)
Hocking, R. R. 1983, Technometrics, 25, 215
Hunter, W. G., & Lamboy, W. F. 1981, Technometrics, 23, 323
Isobe, T., & Feigelson, E. D. 1990, BAAS, 22, 917
Isobe, T., Feigelson, E. D., Akritas, M. J., & Babu, G. J. 1990, ApJ, 364, 104 (IFAB)
Isobe, T., Feigelson, E. D., & Nelson, P. I. 1986, ApJ, 306, 390
Jefferys, W. H., Fitzpatrick, M. J., & McArthur, B. E. 1988a, Celest. Mech., 41, 39
Jefferys, W. H., Fitzpatrick, M. J., McArthur, B. E., & McCartnee, J. E. 1988b, GaussFit (computer program) (available from W. H. Jefferys, Department of Astronomy, University of Texas, Austin, TX 78712)
Jöreskog, K. G. 1973, in Structural Equation Models in the Social Sciences, ed. A. S. Goldberger & O. D. Duncan (New York: Seminar), 85
Jöreskog, K. G., & Sörbom, D. 1984, LISREL VI (computer program) (available from University of Uppsala, Department of Statistics, P.O. Box 513, S-751 20 Uppsala, Sweden)
Jones, T. A. 1979, Math. Geol., 11, 1
Judge, G. G., Hill, R. C., Griffiths, W., Lütkepohl, H., & Lee, T.-C. 1988, Introduction to the Theory and Practice of Econometrics (New York: Wiley)
Kendall, M. G., & Stuart, A. 1977, The Advanced Theory of Statistics, Vol. 2 (4th ed.; London: Griffin)
Krutchkoff, R. G. 1967, Technometrics, 9, 425
Lambert, A., Saleh, A. K., & Sen, P. K. 1985, Commun. Statist. Theor. Meth., 14, 793
LaValley, M., Isobe, T., & Feigelson, E. 1992, in Astronomical Data Analysis Software and Systems, ed. D. Worrall et al. (San Francisco: ASP), in press
Lai, T. L., & Ying, Z. 1992, Stat. Sinica, 2, 17
Lucey, J. R., Bower, R. G., & Ellis, R. S. 1991, MNRAS, 249, 755 (LBE)
Lutz, T. E. 1983, in Statistical Methods in Astronomy, ed. C. Jaschek et al. (Noordwijk: ESA), 179
Lynden-Bell, D., Faber, S. M., Burstein, D., Davies, R. L., Dressler, A., Terlevich, R. J., & Wegner, G. 1988, ApJ, 326, 19
Madansky, A. 1959, J. Am. Stat. Assoc., 54, 173
Maddala, G. S. 1983, Limited-dependent and Qualitative Variables in Econometrics (Cambridge: Cambridge Univ. Press)
Mandel, J. 1984, J. Quality Technol., 16, 1
Martens, H., & Naes, T. 1989, Multivariate Calibration (Chicester: Wiley)
McIntyre, G. A., Brooks, C. Compston, W., & Turek, A. 1966, J. Geophys. Res., 71, 5459
Miller, R. G., & Halpern, J. 1982, Biometrika, 69, 521
Murtagh, F. 1990, in Errors, Bias and Uncertainties in Astronomy, ed. C. Jaschek & F. Murtagh (Cambridge: Cambridge Univ. Press), 385
Neter, J., Wasserman, W., & Kutner, M. H. 1985, Applied Linear Regression Models (2d ed.; Homewood: Irwin)
Osborne, C. 1991, Int. Stat. Rev., 59, 309
Pierce, M. J., & Tully, R. B. 1988, ApJ, 330, 579
Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. 1986, Numerical Recipes: The Art of Scientific Computing (Cambridge: Cambridge Univ. Press)
Reilly, P. M., & Patino-Leal, H. 1981, Technometrics, 23, 221
Riggs, D. S., Guarnieri, J. A., & Addelman, S. 1978, Life Sciences, 22, 1305
Ripley, B. D., & Thompson, M. 1987, Analyst, 112, 377
Rosenblatt, J. R., & Spiegelman, C. H. 1981, Technometrics, 23, 329
Saleh, A. K., & Hassanein, K. M. 1986, Soochow J. Math., 12, 83
Scheffé, H. 1973, Ann. Statist., 1, 1
Schmitt, J. H. 1985, ApJ, 293, 178
Segal, I. E. 1975, Proc. Nat. Acad. Sci., 72, 2473
Sen, P. K. 1968, J. Amer. Stat. Assoc., 63, 1379
———. 1969, Ann. Math. Statist., 40, 1668
Spurrier, J. D., Hewett, J. E., & Lababidi, Z. 1982, Biometrics, 32, 827
Strömberg, G. 1940, ApJ, 92, 156
Teerikorpi, P. 1984, A&A, 141, 407
Tsui, K.-L., Jewell, N. P., & Wu, C. F. 1988, J. Am. Stat. Assoc., 83, 785
Tully, R. B. 1988, Nature, 334, 209
Turner, E. L. 1979, ApJ, 230, 291
Vardi, Y. 1985, Ann. Statist., 13, 118
White, K. J., Haun., S. A., Horsman, N. G., & Wong, S. D. 1988, SHAZAM User's Reference Manual (New York: McGraw-Hill)
Wilcox, R. R. 1987, Brit. J. Math. Statist. Psychol., 40, 80
Working, H., & Hotelling, H. 1929, J. Am. Stat. Assoc., 24 (Suppl.), 73
York, D. 1966, Canadian J. Phys., 44, 1079
———. 1967, Earth Planet. Sci. Lett., 2, 479