SOME EFFECTS OF ERRORS OF MEASUREMENT ON LINEAR REGRESSION

W. G. COCHRAN
HARVARD UNIVERSITY

1. Introduction

I assume a bivariate distribution of pairs (y, X) in which y has a linear regression on X

$$y = \beta_0 + \beta_1 X + e, \tag{1.1}$$

where e, X are independently distributed and $E(e|X) = 0$. However, the measurement of X is subject to error. Thus we actually observe pairs (y, x), with $x = X + h$, where h is a random variable representing the error of measurement.

Given a random sample of pairs (y, x), previous writers have discussed various approaches to the problem of making inferences about the line $\beta_0 + \beta_1 X$, sometimes called the *structural* relation between y and X. In the present context this line might be called "the regression of y on the correct X" to distinguish it from "the regression of y on the fallible x." An obviously relevant question is: under assumption (1.1), what is the nature of the regression of y on x?

Lindley [5] gave the necessary and sufficient conditions that the regression of y on the fallible x be linear in the narrow sense. This means that $E(y|x)$ is linear in x, or equivalently that

$$y = \beta_0' + \beta_1' x + e', \tag{1.2}$$

where $E(e'|x) = 0$. This definition does not require that e' and x be independently distributed. Lindley's proof assumes that the error of measurement h is distributed independently of X. His necessary and sufficient conditions are that Fisher's cumulant function (logarithm of the characteristic function) of h be a multiple of that of X. Roughly speaking, this implies that h and X belong to the same class of distributions. Thus if X is distributed as $\chi^2\sigma^2$, so is h, though the degrees of freedom can differ: if X is normal, h must be normal.

Several writers have discussed the corresponding necessary and sufficient conditions if we demand in addition that the residual e' in (1.2) be distributed independently of x. In particular, Fix [3] showed that if the second moment of

This work was supported by the Office of Naval Research through Contract N00014-56A-0298-0017, NR-042-097 with the Department of Statistics, Harvard University.

either X or h exists and X, h are independent as before, the conditions for linearity of regression in this fuller sense are that both X and h be normally distributed.

Thanks partly to computers, numerous regression studies are being done nowadays, particularly in the social sciences and medicine, in which all the variables are difficult to measure, and therefore are presumably measured with sizeable errors. In thinking about mathematical models appropriate to such uses, it seems clear that the forces which determine the nature of the distribution of h (the imperfections of the measuring instrument or process) are quite different from those that determine the nature of the distribution of the correct X. Consequently, my opinion is that in such applications even the Lindley conditions will not be satisfied, except perhaps by a fluke or as an approximation (for example, the cumulants of X and h might be similar in the sense that both distributions are close to normal).

This paper considers the regression of y on x when Lindley's conditions are not satisfied. There are at least two reasons for interest in this regression. The objective may be to obtain a consistent estimate of β_1 for purposes of interpretation or adjustment by covariance (Lord [6]). Secondly, the purpose may be to predict y from the fallible x by the regression technique in which case the shape of this regression is relevant.

The strategy used here is first to construct a straight line relation between y and the fallible x which may be called the *linear component* of the regression of y on x. This is the line that we are estimating, in some sense, when we compute a sample linear regression of y on x.

The paper then takes a look at the question: what is the nature of the departure from linearity when Lindley's conditions are not satisfied? In particular, does the linear component dominate? If it does, then as Kendall [4] remarks, "A slight departure from linearity will sometimes allow the ordinary theory to be used as an approximation." I have been unable to obtain any general results that are exact, but something can be learned by a combination of an approach *via* moments and the working out of some easy particular cases. These suggest, fortunately, that the linear component often dominates, even with measurements of rather poor reliability, but the issue needs more thorough investigation by someone with greater mathematical power.

2. The linear component

As stated, a linear regression of y on the correct X (in the fullest sense) is assumed, namely,

$$y = \beta_0 + \beta_1 X + e, \tag{2.1}$$

where e, X are independently distributed and $E(e|X) = 0$. I also assume X scaled so that $E(X) = 0$.

As regards the error of measurement h, Lindley's result requires h, X to be independently distributed, but this assumption limits the range of applications of the result. Some measuring instruments or methods underestimate high values of X and overestimate low values. I have been unable to obtain conditions analogous to Lindley's when X, h follow a general bivariate distribution $\phi(X, h)$ but there is no difficulty in obtaining the linear component in this case. Denote $E(h)$ by μ_h, since measurements may be biased. It is assumed, as seems reasonable for most applications, that h and hence x are distributed independently of e. Hence the regression of y on x is, from (2.1),

$$E(y|x) = \beta_0 + \beta_1 E(X|x) = \beta_0 + \beta_1 R(x), \tag{2.2}$$

say.

Thus we need to find $R(x)$. Let $\phi(X, h)$ be the joint frequency function of X, h. The marginal distribution of the fallible x is

$$\psi(x) = \int \phi(X, x - X)\, dX, \tag{2.3}$$

while $R(x) = E(X|x)$ satisfies the equation

$$R(x)\psi(x) = \int X\phi(X, x - X)\, dX. \tag{2.4}$$

The linear component of $R(x)$ can be defined by fitting the straight line $L(x) = C_0 + C_1 x$ to $R(x)$ by the population analogue of the method of least squares. That is, we choose C_0, C_1 to minimize

$$\int \{R(x) - C_0 - C_1 x\}^2 \psi(x)\, dx. \tag{2.5}$$

Clearly,

$$\int R(x)\psi(x)\, dx = \iint X\phi(X, h)\, dX dh = \mu_X = 0, \tag{2.6}$$

$$\int xR(x)\psi(x)\, dx = \iint (X^2 + Xh)\phi(X, h)\, dX dh = \sigma_X^2 + \sigma_{Xh}, \tag{2.7}$$

where σ_{Xh} is the population covariance of X and h. Hence the normal equations for C_0 and C_1 give

$$C_0 = -C_1\mu_h, \quad C_1 = (\sigma_X^2 + \sigma_{Xh})/(\sigma_X^2 + \sigma_h^2 + 2\sigma_{Xh}). \tag{2.8}$$

From (2.2), the linear component of the regression of y on x is $L(x) = \beta_0 + \beta_1(C_0 + C_1 x)$. If we write this $\beta_0' + \beta_1' x$, we have

$$\beta_0' = \beta_0 - \beta_1 C_1 \mu_h, \quad \beta_1' = \beta_1 C_1 = \beta_1(\sigma_X^2 + \sigma_{Xh})/(\sigma_X^2 + \sigma_h^2 + 2\sigma_{Xh}). \tag{2.9}$$

Incidentally, expressions (2.9) for β'_0, β'_1 can be obtained directly by noting that our procedure is equivalent to defining $(\beta'_0 + \beta'_1 x)$ as the linear component of the regression of y on x if we write $y = \beta'_0 + \beta'_1 x + e'$ and determine β'_0, β'_1 so that the residuals e' satisfy the conditions

$$E(e') = 0, \qquad \text{Cov}\,(e', x) = 0. \tag{2.10}$$

Formulas (2.9) for β'_0 and β'_1 agree with the well-known elementary results in the literature, usually obtained on the assumption that X, h follow a bivariate normal. Bias in the measurements affects the intercept β'_0 but not the slope β'_1. If h and X are uncorrelated, $\beta'_1 = \beta_1 \sigma_X^2/\sigma_x^2$, the factor σ_X^2/σ_x^2 being often called the *reliability* of the measurement x. For given σ_X^2, σ_h^2, positive correlation of the errors with X makes the underestimation of the slope worse, while negative correlation alleviates it if $\sigma_h^2 < \sigma_X^2$. In the Berkson case [1], the investigator plans to apply preselected amounts x of some agent or treatment in a laboratory experiment, but owing to errors in measuring out this amount, the amount X actually applied is different. Here, Cov $(x, h) = 0$, so that $\sigma_{Xh} = -\sigma_h^2$ and (2.9) shows that $\beta'_1 = \beta_1$. This situation also applies when large samples are grouped by their values of X into classes to facilitate the calculation of regression on a desk machine, provided that x is taken as the *mean* of X within each class. The common practice is of course to take x as the midpoint of the class. This makes Cov (x, h) slightly positive for most unimodal distributions of X, so that some residual inconsistency in β'_1 as an estimate of β_1 remains, though the inconsistency is in general trivial if at least ten classes are used.

Suppose now that y is also subject to an error of measurement d. If Y represents the correct value of y, we may rewrite the original model (1.1) as

$$Y = \beta_0 + \beta_1 X + e, \qquad y = Y + d. \tag{2.11}$$

Hence

$$E(y|x) = \beta_0 + \beta_1 R(x) + E(d|x). \tag{2.12}$$

If errors in y are independent of Y, X and h, then $E(d|x) = \mu_d$, the amount of bias in d, and we get the old result that such errors in y do not affect the slope of the regression line. If d is correlated with Y, the choice of an appropriate model requires care. Specification of the joint frequency functions $\phi(X, h)$, $\theta(Y, d)$ is not enough to determine $E(d|x)$; we need to know the relation between d and h. The following might serve for applications in which the process by which y is measured is independent of that by which x is measured. Noting that $E(X) = 0$, $E(Y) = \beta_0$, write

$$d = \mu_d + \frac{\sigma_{Yd}}{\sigma_Y^2}(Y - \beta_0) + d', \qquad h = \mu_h + \frac{\sigma_{Xh}}{\sigma_X^2} X + h', \tag{2.13}$$

where h', d', with zero means, are assumed independent of each other and of X, Y, and e. This model does not imply that d and h are independent, since

$$\text{Cov}\,(dh) = \frac{\sigma_{Yd}\sigma_{Xh}\sigma_{XY}}{\sigma_Y^2\sigma_X^2}, \tag{2.14}$$

but this correlation arises only as a consequence of the X, Y correlation.

In some applications there may be further correlation between d and h because the measuring processes are not independent. For instance, an individual pair (x, y) might be estimates of a town population five years apart, where the municipal statisticians use the same techniques in a town, the technique varying from town to town. In general, it will obviously be difficult to know which model to pursue, and to get data for verification of a model.

If (2.13) holds,

$$E(d|x) = \mu_d + \frac{\sigma_{Yd}}{\sigma_Y^2}\beta_1 R(x). \tag{2.15}$$

From (2.8) and (2.12), we obtain for the linear component of the regression of y on x,

$$\left[\beta_0 + \mu_d - \beta_1 C_1 \mu_h \frac{(\sigma_Y^2 + \sigma_{Yd})}{\sigma_Y^2}\right] + \beta_1 x \frac{(\sigma_Y^2 + \sigma_{Yd})(\sigma_X^2 + \sigma_{Xh})}{\sigma_Y^2(\sigma_X^2 + 2\sigma_{Xh} + \sigma_h^2)}. \tag{2.16}$$

As is obvious from graphical considerations, errors in y that are positively correlated with Y tend to increase the absolute value of β_1', whereas errors in x have the opposite effect. With errors in both y and x, β_1' may be either greater or less than β_1.

The method of obtaining the linear component extends naturally to a multiple linear regression of y on $x_1, x_2, \cdots, x_k$. Even when the errors of measurement h_i are independent of X_i and of each other—the simplest case—β_i' is a linear function of all β_j whose corresponding x_j are subject to errors of measurement (Cochran [2]). When the h_i and X_i are correlated, we again meet the problem of specifying the nature of the correlation between h_i and h_j.

One objective in working out the relations between β_i' and the β_j is as a possible means of estimating the coefficients β_j of the structural regression by using data from supplementary studies of the errors of measurement. With errors in more than one variate, however, the algebraic results suggest that the information needed about errors of measurement is more than we are likely to be able to obtain.

3. Polynomial approach by moments

Now consider the nature of $R(x)$ with errors in x only when Lindley's conditions are not satisfied. Like Lindley, I assume h and X independent, with $\mu_h = 0$. I first chose some simple forms for the frequency functions $f(X)$ of X and $g(h)$ of h for which $R(x)$ can be worked out exactly in closed form. Examples are the χ distribution with a small number of degrees of freedom, the normal, the uniform, and the exponential types like e^{-X}, with $X > 0$, or $\frac{1}{2}e^{-|x|}$, with $-\infty < X < \infty$. Inspection of a few cases indicated that if either X or h follows a

skew distribution, the departure of $R(x)$ from linearity, in a region around the mean of x, is of the simple type that can be approximated by a quadratic curve (an example will be given in Section 4).

If, however, both h and X are symmetrically distributed about their means 0, then $\psi(x)$ is also symmetrical and $R(-x) = -R(x)$, which suggests a cubic approximation with a zero quadratic term. The equations of the approximating quadratic or cubic, by the least squares method, are obtained easily from the low moments of the distributions of h and X. (In the symmetric case it is possible that for some frequency functions a quadratic approximation in $|x|$, with reversal of sign when x is negative, might do better than the cubic, but the fitting requires calculation of some incomplete moments and this has not been pursued.)

In fitting a polynomial approximation of degree p to $R(x)$, we choose the coefficients C_i to minimize

$$\int \{R(x) - \sum_{i=0}^{p} C_i x^i\}^2 \psi(x)\, dx. \tag{3.1}$$

The rth normal equation is

$$\sum_{i=0}^{p} C_i \mu_{r+i,x} = \int\int X(X + h)^r f(X) g(h)\, dX dh, \tag{3.2}$$

where the μ denote moments about the mean.

Here we are fitting only the simplest nonlinear approximations, say $Q(x)$.

Case 1. *X or h skew.*

$$Q(x) = C_1 x + C_2(x^2 - \mu_{2x}), \tag{3.3}$$

where

$$\Delta = \mu_{4x}\mu_{2x} - \mu_{3x}^2 - \mu_{2x}^3, \tag{3.4}$$

$$C_1 = (\mu_{4x}\mu_{2X} - \mu_{3x}\mu_{3X} - \mu_{2x}^2\mu_{2X})\Delta^{-1}, \tag{3.5}$$

$$C_2 = (\mu_{2x}\mu_{3X} - \mu_{3x}\mu_{2X})\Delta^{-1}. \tag{3.6}$$

Case 2. *X and h both symmetrical.*

$$C(x) = c_1 x + c_3 x^3, \tag{3.7}$$

where

$$\Delta = \mu_{6x}\mu_{2x} - \mu_{4x}^2, \tag{3.8}$$

$$c_1 = (\mu_{6x}\mu_{2X} - \mu_{4x}\mu_{4X} - 3\mu_{4x}\mu_{2X}\mu_{2h})\Delta^{-1}, \tag{3.9}$$

$$c_3 = (\mu_{4x}\mu_{2X} - \mu_{2x}\mu_{4X} - 3\mu_{2x}\mu_{2X}\mu_{2h})\Delta^{-1}. \tag{3.10}$$

There is obvious interest in seeing how well the quadratic and cubic approximations fit $R(x)$. The reduction in the variance of $R(x)$ due to these approximations can be obtained from the normal equations by the usual analysis of variance rule of multiplying the solutions C_i or c_i by the right sides of the normal

equations. But I have been unable to obtain an exact expression for the variance of $R(x)$ which does not involve computing an integral, so that I do not have a general result for the closeness of fit in this case, though it can be obtained by numerical integration in specific examples, as discussed in Section 5.

This approach extends also to correlated errors, the expressions for the C_i involving joint moments of h and X which are easily obtained.

4. Two examples

As an example of the quadratic approximation I take

$$f(X) = Xe^{-X^2/2}, \qquad X > 0, \qquad g(h) = N(0, \sigma^2). \tag{4.1}$$

Thus X is skew, essentially a χ variate with two degrees of freedom, mean $\sqrt{\pi/2}$ and variance $(2 - \pi/2)$, about 0.429, while h is normal. The reliability of the measurement is $(4 - \pi)/(4 - \pi + 2\sigma^2)$, so that measurements with different degrees of reliability from 92 per cent to 54 per cent are represented by taking $\sigma = 0.2(0.1)0.6$. Measurements of lower reliability have been reported, but this range should cover the great majority of applications. Reliability of 50 per cent is far from impressive: the measurement error has as big a variance as the correct measurement. It is not claimed that this example corresponds to any actual situation in practice: although X is essentially positive, x is not.

For this example, $R(x)$ works out as

$$R(x) = \frac{\sigma}{(1 + \sigma^2)^{1/2}} \left\{ \frac{(u^2 + 1)P(u) + uz(u)}{uP(u) + z(u)} \right\} \tag{4.2}$$

where $u = x/\sigma(1 + \sigma^2)^{1/2}$, $z(u)$, and $P(u)$ are the ordinate and cumulative of $N(0, 1)$. (In this example X was not scaled so that $E(X) = 0$). Figure 1 presents points on $R(x)$ for $\sigma = 0.3$ and $\sigma = 0.6$, plus the line $R(x) = x$ that would apply if there were no error of measurement.

As suggested, the major departure from linearity of the points in Figure 1 near μ_x is a simple curvature that is well represented by a quadratic curve, though this quadratic would do very badly at points far away from μ_x. For instance, with x and u becoming large and positive, $P(u)$ tends to one and $z(u)$ to zero, so that $R(x)$ behaves like $x/(1 + \sigma^2)$, while for x negative $R(x)$ tends to become asymptotic to the origin. However, the approximation errors in the quadratic $Q(x)$ receive very little weight at these points, since they are far enough away from the mean of x so that $\psi(x)$ is tiny. The mean square error (MSE) of the quadratic approximation, the integral of $[Q(x) - R(x)]^2\psi(x)\,dx$, is only around 0.0003 to 0.0007. The MSE of the linear approximation has the following values.

σ	.2	.3	.4	.5	.6
MSE$[L(x)]$	.0010	.0025	.0041	.0052	.0069

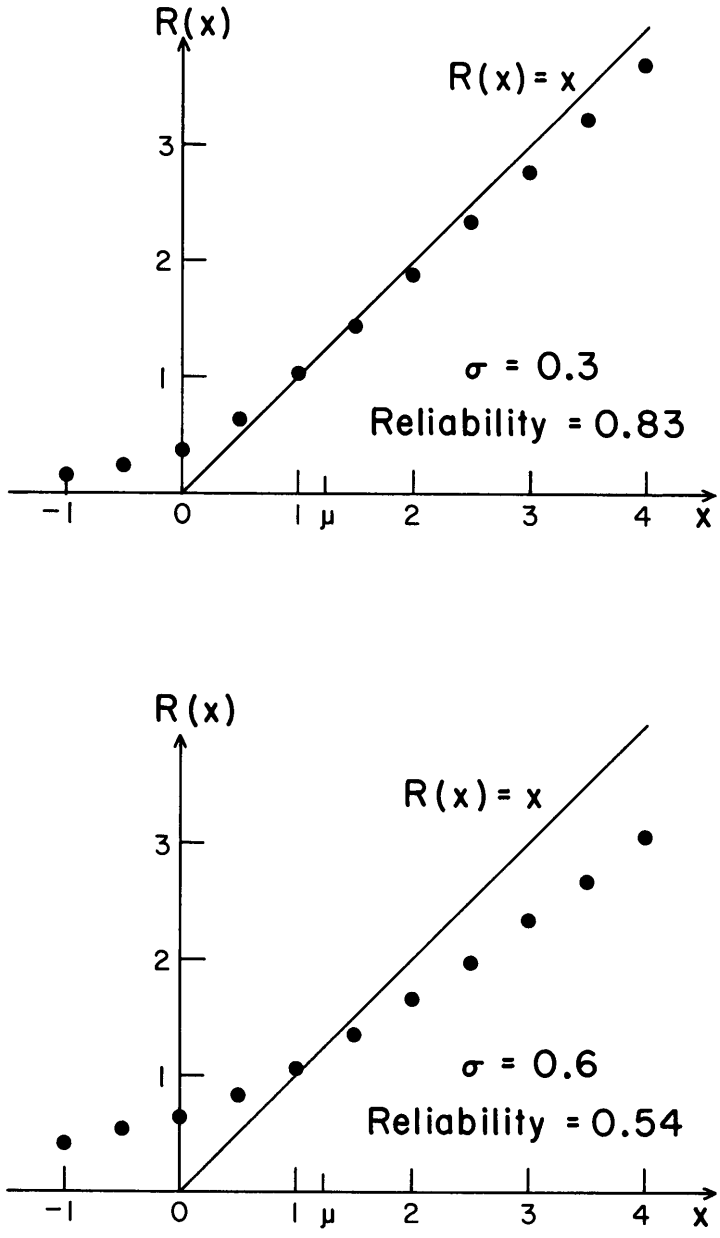

FIGURE 1
Regression $R(x)$ of X (correct) on x (fallible) measurement.
X is χ (2 d.f.), error of measurement is $N(0, \sigma^2)$.

The example of the symmetrical case has $X = N(0, 1)$, while the measurement error h is uniform between $-L$ and L. The reliability is $3/(3 + L^2)$ and with $L = 0.5(0.25)1.5$, it varies between 93 per cent and 57 per cent. Figure 2 shows $R(x)$ for $L = 1$, $L = 1.5$, with reliabilites 0.75 and 0.57. In this example

$$R(x) = [z(x - L) - z(x + L)][C(x + L) - C(x - L)]^{-1}. \tag{4.3}$$

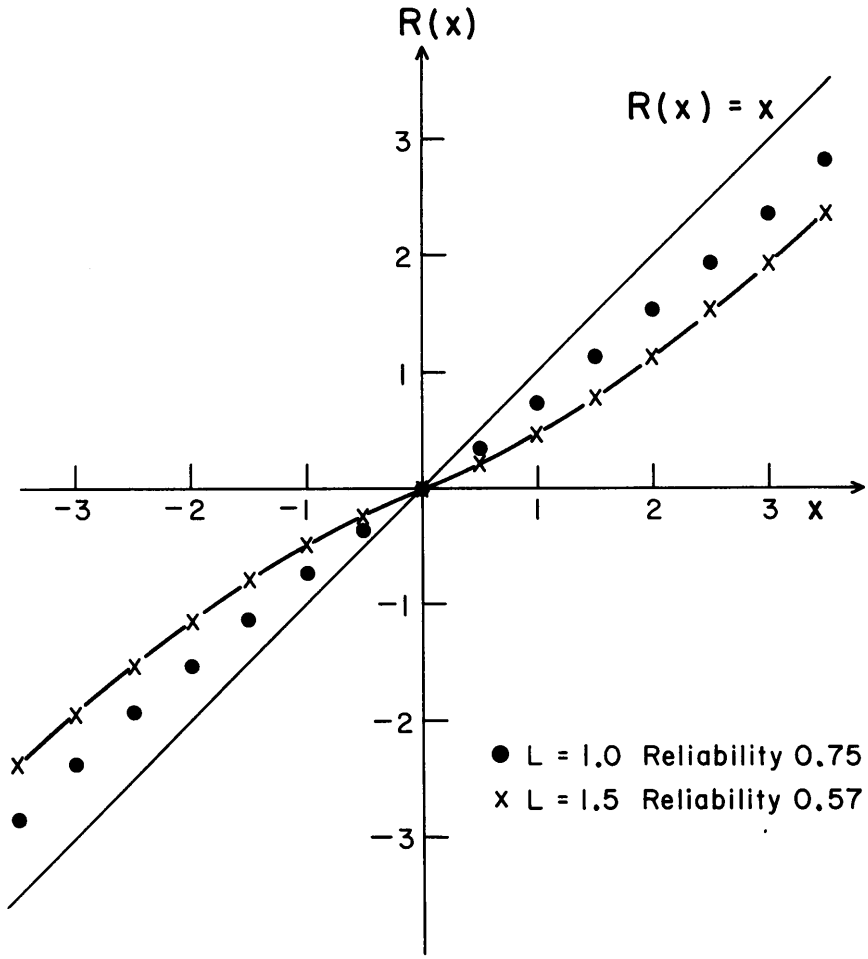

FIGURE 2
Regression of $R(x)$ of X (correct) on x (fallible) measurement.
X is $N(0, 1)$, error of measurement is uniform $(-L, L)$.

As x moves far away from its mean at 0, $R(x)$ tends to become $(x - L)$, again linear in x rather than cubic. The MSE of the cubic approximation is very small, the greatest value in the cases worked being 0.00054 at $L = 1.5$. For the linear approximation, the MSE are as follows.

L	0.5	0.75	1	1.25	1.5
MSE$[L(x)]$	$.0^45$	$.0^316$	.0011	.0046	.0116

5. Adequacy of the linear approximation

In considering the adequacy of the linear approximation, I am not concerned with applications in which the aim is to estimate the structural relationship, but only with those in which (i) the objective is to predict y from x, in which case the usual advice is to utilize the observed regression of y on x, or (ii) to test the null hypothesis $\beta'_1 = 0$ by a t test, as a means of testing the null hypothesis $\beta_1 = 0$.

First, as Kendall [4] reminded us, Lindley's type of linearity falls short of the assumptions of independence of the residual e' and x and normality of e' that are needed for use of the standard regression formulas. The situation resembles that in the standard texts on sample surveys, in which attempts have been made to construct a theory of linear regression estimators without allotting any specific structure to a relation between y and x except that all values of y and x in the population are bounded. It is known in such cases that the usual sample estimate $\hat{\beta}'_1$ of β'_1 is biased, the leading term in the bias being $-E(e'x^3)/n\sigma_x^4$, where n is the sample size. The usual formula for the variance of $\hat{\beta}'_1$ holds as a first approximation, but it too has a bias that becomes negligible in large samples, while the numerator and denominator of the t test of $\hat{\beta}'_1$ only become independent asymptotically. At best we can say that the usual methods apply asymptotically.

As regards the effect of errors in x on the precision of regression estimates of y, the most important quantity is the variance of the residuals from the regression of y on x, or from approximations to this regression that we consider using. It is worth starting with the simplest case in which X, h are normal and independent, so that the regression of y on x is linear in both senses. Here $\beta'_1 = \beta_1\sigma_X^2/\sigma_x^2 = G\beta_1$, where G is the coefficient of reliability. Further, since

$$\sigma_y^2 = \sigma_e^2 + \beta_1^2\sigma_X^2 = \sigma_{e'}^2 + \beta_1'^2\sigma_x^2 = \sigma_{e'}^2 + G\beta_1^2\sigma_X^2, \tag{5.1}$$

we have

$$\sigma_{e'}^2 = \sigma_e^2 + (1 - G)\beta_1^2\sigma_X^2 = (1 - \rho^2)\sigma_y^2 + (1 - G)\rho^2\sigma_y^2. \tag{5.2}$$

This is the familiar result that, even when there is no problem about nonlinearity, errors in x increase the variance of the deviations from the linear prediction model by $(1 - G)\rho^2\sigma_y^2$, an increase that hurts most, for given G, when the prediction formula is very good (ρ high).

Returning to the assumptions of this paper, suppose y is predicted from its regression on x. Since

$$y = \beta_0 + \beta_1 X + e = \beta_0 + \beta_1 R(x) + e', \tag{5.3}$$

$$\sigma_{e'}^2 = \sigma_e^2 + \beta_1^2(\sigma_X^2 - \sigma_R^2) = \sigma_e^2 + \beta_1^2\sigma_X^2(1 - \sigma_R^2/\sigma_X^2). \tag{5.4}$$

In the examples that I have worked, σ_R^2/σ_X^2 comes numerically to within one or two percentage points of $G = \sigma_X^2/\sigma_x^2$. Consequently, if the population regression of y on x is used for prediction, the loss of precision due to errors in x is about what we expect from the value of the reliability coefficient G.

To pass to the additional loss of precision if $L(x)$ instead of $R(x)$ is used in the prediction formula, write

$$\beta_0 + \beta_1 R(x) + e' = \beta_0 + \beta_1 L(x) + e'', \tag{5.5}$$

giving

$$\sigma_{e''}^2 = \sigma_{e'}^2 + \beta_1^2(\sigma_R^2 - \sigma_L^2) = \sigma_{e'}^2 + \beta_1^2 \text{ MSE } (L) \tag{5.6}$$

since the method of fitting L makes Cov $(L, R - L) = 0$. From the little tables of MSE (L) in Section 4, the highest ratios of MSE (L) to σ_X^2, which occur when reliability is lowest, are around 0.01 to 0.02. If these examples typify what happens in applications, it appears that errors in x create an increase of around $(1 - G)\rho^2\sigma_y^2$ in the variance of residuals, but that even when reliability is low, the additional increase in this variance due to use of $L(x)$ instead of $R(x)$ is unimportant.

Since e' and x are not independent, a look at the conditional distribution of e' for fixed x may be worthwhile. From (5.2),

$$e' = e + \beta_1[X - R(x)]. \tag{5.7}$$

By our hypothesis, the term e is independent of X, h, and hence x, so that this term has the same shape of distribution and variance in all arrays with x fixed. The second term is determined by the distribution of X for fixed x. In Example 1 this distribution works out as

$$f(X|x) = \frac{(1 + \sigma^2)}{\sigma^2\sqrt{2\pi}} \frac{1}{K(x)} X \exp\left\{-\frac{1}{2}\frac{(1 + \sigma^2)}{\sigma^2}\left[X - \frac{x}{1 + \sigma^2}\right]^2\right\} \tag{5.8}$$

where, with $\mu = x/\sigma(1 + \sigma^2)^{1/2}$,

$$K(x) = z(u) + uP(u). \tag{5.9}$$

This distribution changes in shape as x varies; its variance increases with x, the increase being small when the reliability G is high but more marked when G is lower.

In Example 2, $f(X|x)$ is simply the incomplete normal

$$f(X|x) = \frac{1}{\sqrt{2\pi}} \frac{1}{K(x)} \exp\{-\tfrac{1}{2}X^2\}; \qquad x - L \leqq X \leqq x + L \tag{5.10}$$

with $K(x) = P(x + L) - P(x - L)$. When x is at its mean 0, this is symmetrical with its maximum variance, but changes from negative to positive skewness as x changes from negative to positive.

Thus the distribution of e' is a compound of two independent distributions, one unchanging, the other changing with x. As we have seen, the average variances of the two components are $\sigma_y^2(1 - \rho^2)$ and approximately $\rho^2\sigma_y^2(1 - G)$. With G high and ρ modest, the unchanging component should dominate, and the assumption that the distribution of e' is the same for different x may be a reasonable approximation, but with say $\rho = 0.8$, $G = 0.5$, the two components have variances $0.36\sigma_y^2$ and $0.32\sigma_y^2$ and are about equally important.

6. Summary and discussion

This paper deals with applications in which the standard linear regression model $y = \beta_0 + \beta_1 X + e$, with e, X independent and $E(e) = 0$, is assumed to apply to a bivariate sample of pairs (y, X). However, owing to difficulties in measuring the X values, we actually have a bivariate sample (y, x) where $x = X + h$, h being an error of measurement. My opinion is that in applications even Lindley's conditions for linearity of the regression of y on x in the narrow sense will not in general be satisfied.

Facing this situation we can define a linear relation $y = \beta'_0 + \beta'_1 x$ that may be called the linear component of the regression of y on x. The elementary results in the literature for the relations between β'_0, β'_1 and β_0, β_1 (usually derived on the assumption that h, X are normally distributed) hold for this linear component. The linear component can be obtained when h and X are correlated, whereas Lindley assumes h, X independent, and extends to errors in y also and to multiple linear regression, subject to problems about specification of the nature of correlated errors.

The next step was to work out the exact regression of y on x for a number of specific examples in which this regression has a closed form. These suggest that the departure from linearity can be approximated by a quadratic in x if either h or X is skew and by a cubic in x if h, X are symmetrical. The equations of the approximating quadratic or cubic are easily obtainable from the lower moments of the distributions of h and X. Further, in these examples the linear component dominates, in the sense that the mean square deviation of y from the linear component is only slightly larger than that from the exact regression of y on x, even for measurements of reliability not much more than 50 per cent.

A further result that holds well in these examples is of interest when the fallible x is used to predict y. When h and X are independent and normal, so that the regression of y on x is linear, it is known that the residual variance from the regression of y on x exceeds that for the regression of y on X by $(1 - G)\rho^2\sigma_y^2$, where G is the coefficient of reliability of x. This result remains a good approximation when X, h are independent but have different distributions so that Lindley's conditions do not hold.

Unfortunately the results here are only suggestive and leave unanswered the important questions. For example, (i) what is the analogous result to Lindley's when h and X are not independent? Some measuring instruments have the property of underestimating high values of X and overestimating low values, and *vice versa*. (ii) How far can the moments approach be trusted? Are there distributions for which the departure from linearity is more complex than a quadratic or cubic? (iii) Lindley's conditions guarantee linearity only in the narrow sense; the deviations e' from the regression of y on x are not independent of x. There is reason to believe that in this case linear regression theory can be used asymptotically in large samples, but more needs to be known about the practical importance of the disturbances present in small samples.

REFERENCES

[1] J. BERKSON, "Are there two regressions?," *J. Amer. Statist. Assoc.*, Vol. 45 (1950), pp. 164–180.

[2] W. G. COCHRAN, "Errors of measurement in statistics," *Technometrics*, Vol. 10 (1968), pp. 637–666.

[3] E. FIX, "Distributions which lead to linear regressions," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1949, pp. 79–91.

[4] M. G. KENDALL, "Regression, structure and functional relationship." *Biometrika*, Vol. 38 (1951), pp. 11–25.

[5] D. V. LINDLEY, "Regression lines and the linear functional relationship," *J. Roy. Statist. Soc. Ser. B*, Vol. 9 (1947), pp. 218–244.

[6] F. LORD, "Large-sample covariance analysis when the control variable is fallible," *J. Amer. Statist. Assoc.*, Vol. 55 (1960), pp. 307–321.