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

(1.1) $$y = \beta_0 + \beta_1 X + e,$$

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

(1.2) $$y = \beta'_0 + \beta'_1 x + e',$$

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.