ERRORS IN VARIABLES FOR THE SIMPLE LINEAR MODEL: THE EFFECTS OF CORRELATED ERRORS OF MEASUREMENT ON INTERVAL ESTIMATION AND HYPOTHESIS TESTING

John J. Chai and George Frankfurter, Syracuse University

1. INTRODUCTION

1.1 Purpose and scope: It is well known that the Ordinary Least Squares Estimator (OLSE) of the regression coefficient for the simple linear regression model

$$Y = \alpha + \beta X + \varepsilon$$

is biased and not consistent when both variables are subject to errors of measurement [9]. A number of studies in this area largely deal with simple models [e.g., 1, 3, 6, 8, 10]. In these studies each of the "true" values and the corresponding error terms are assumed to be uncorrelated. Also, the two error terms are assumed to be uncorrelated. There were some exceptions however. Chai [1] and Cochran [3] disucss some effects of correlation between the two error terms and between the "true" independent variable (X) and its errors of measurement. Although no empirical results have been published, some have mentioned possible cases where correlation between the "true" independent variable and its measurement errors may exist.

The purpose of this paper is to examine, for a range of different values of ρ_{XY} and other relevant parameters, the bias of the Ordinary Least Squares Estimator of β due to

(1) Correlation between the errors of measurement of the "true" values (i.e., ρ_{de}),

(2) Correlation between each of the errors of measurement and the respective "true" values (i.e., ρ_{Xd}, ρ_{Xe}, ρ_{Yd}, ρ_{Ye}).

Further, the paper shows the effect of the bias of the OLSE of β on the confidence interval estimation for β and the hypothesis testing for $\beta=0$ v.s. $\beta\neq 0$.

1.2 The rationale for this study: In econometric research it quite often happens that there would be certain relationships between errors of measurement and the "true" variables and between measurement errors themselves. For example, in measuring the values of housing, the error of measuring expensive units may be positively correlated with the "true" values of the units, because there exists a tendency to undervalue housing in ghetto areas and overestimate property value in suburban areas. Another case may be measurement of family income. The error of measurement might be negatively correlated with the "true" family income, because of the tendency of under-reporting of higher income and of over-reporting lower income.

Still another group of examples could be found in time series data. If the independent variable is a function of time, and if the errors of measurement for the "true" value for the earlier period are greater (because of lack of memory, for example) than the ones for the recent measurement, then one would expect a correlation not only between the "true" value and the error of measurement, but also a correlation between errors over successive time periods.

In the case of a stock market index it may happen that the majority of the securities in the index over-react to the general trend of the market. They go up more on an "up-market" and go down more on a "down-market" than the average movement of the universe of all securities.

All these socio-economic interrelationships between the "true" values of the two variables and between their respective errors of measurement warrant a careful examination of the effects of various errors of measurement on the OLSE of β. In section 2 the mathematical model, the bias, and the bias relative to the true β are presented. A discussion on the effect of the bias on the OLSE is presented in Section 3.

2. THE MODEL

2.1 Bias: Let Y and X respectively be the "true" values for a population element and let the relationship between Y and X be given by a linear regression model

$$Y = \alpha + \beta X + \varepsilon \qquad (1)$$

where ε is a random residual of the regression. We assume that both X and Y are subject to errors of measurement and let the values actually obser observed be:

$$x = X + d \qquad (2)$$

$$y = Y + e \qquad (3)$$

where d and e represent measurement errors. Substituting (2) and (3) into (1) results in

$$y = \alpha + \beta x + u \qquad (4)$$

$$\text{where } u = -\beta d + e + \varepsilon \qquad (5)$$

we assume that X, Y, d, and e jointly follow a multi-variate normal process with mean vector $[\mu_X, \mu_Y, \mu_d, \mu_e]$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{Xd} & \sigma_{Xe} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YD} & \sigma_{Ye} \\ \sigma_{Xd} & \sigma_{Yd} & \sigma_d^2 & \sigma_{de} \\ \sigma_{Xe} & \sigma_{Ye} & \sigma_{de} & \sigma_e^2 \end{pmatrix}$$

Hence, the regression given by (4) above is linear according to Lindley [11].

Without loss of generality we assume that all means are put equal to zero (α=0) for the purpose of deriving the bias of the OLSE. Let the OLSE of β be denoted by b. Then $b = \Sigma xy/\Sigma x^2$ and for the regression of y on x, $E(b) = E\{xE(y|x)/\Sigma x^2\}$. Furthermore, the regression of y on x is linear so if we write $E(y|x) = \lambda x$ we have $E(b) = \lambda$. λ is found, therefore, by evaluating $E(y|x)/x$. But first, evaluate $E(y|x)$.

$E(y|x) = \beta x + E(u|x) = \beta x + E(e|x) - \beta E(d|x) + E(\varepsilon|x)$. Since $\sigma_{X\varepsilon} = 0$ and $\sigma_{d\varepsilon} = 0$, we have

$$E(y|x) = \beta x + E(e|x) - \beta E(d|x). \qquad (6)$$

We now evaluate the last two terms of the right-hand side of equation (6) each divided by x as follows:

$$\frac{E(d|x)}{x} = \frac{E(dx)}{E(x^2)} = \frac{E\{d(X+d)\}}{E(X+d)^2} = \frac{\sigma_{Xd} + \sigma_d^2}{\sigma_X + 2\sigma_{Xd} + \sigma_d^2}$$

$$\frac{E(e|x)}{x} = \frac{E(ex)}{E(x^2)} = \frac{E\{e(X+d)\}}{E(X+d)^2} = \frac{\sigma_{Xe} + \sigma_{de}}{\sigma_X^2 + 2\sigma_{Xd} + \sigma_d^2}$$

Thus, by substitution

$$\lambda = \beta + \frac{\sigma_{Xe} + \sigma_{de} - \beta(\sigma_{Xd} + \sigma_d^2)}{\sigma_X^2 + 2\sigma_{Xd} + \sigma_d^2} \qquad (7)$$

The bias is therefore the second term on the right-hand side of (7). Let the bias be denoted by B_b. Rewriting it in terms of correlation coefficients ρ_{Xd}, ρ_{Xe}, and ρ_{de}, we obtain

$$B_b = \frac{\rho_{Xe}\sigma_X\sigma_e + \rho_{de}\sigma_d\sigma_e - \beta(\rho_{Xd}\sigma_X\sigma_d + \sigma_d^2)}{\sigma_X^2 + 2\rho_{Xd}\sigma_X\sigma_d + \sigma_d^2} \qquad (8)$$

We define the relative biases as follows:

(1) Bias relative to β: B_b/β (9)

(2) Bias relative to σ_b: B_b/σ_b (10)

where σ_b is the standard error of b, i.e.,

$$\sigma_b = \sigma_u / \sqrt{\Sigma(x-\bar{x})^2}$$

and from (4) and (5)

$$\sigma_u^2 = \Sigma(y - \alpha + \beta x)^2 = \Sigma(\varepsilon + e - \beta d)^2$$
$$= \sigma_\varepsilon^2 + \sigma_e^2 + \beta^2\sigma_d^2 - \beta\sigma_{de} \qquad (11)$$

2.2 Effect of bias on confidence interval for β: Assuming that we have a multivariate normal distribution of X, Y, d, and e, the OLSE of β, b is normal. Let b_L and b_U respectively be defined as follows:

$$b_L = \beta - |z|\,\sigma_b$$

$$b_U = \beta + |z|\sigma_b$$

where

$$z = \{b - E(b)\}/\sigma_b$$

Now let Z_L and Z_U respectively be the standardized values of b_L and b_U. Then

$$Z_L = -|z| - B_b/\sigma_b$$

$$Z_U = |z| - B_b/\sigma_b$$

The z indicates the desired level of confidence and Z_L and Z_U indicate the actual level of confidence realized.

2.3 Effect of bias on testing of hypotheses β=0 v.s. $\beta\neq 0$: In testing of the null hypothesis H_0: β=0 against the alternative H_1: $\beta\neq 0$ with Type I error controlled, the critical values (action limits) in terms of standardized values, A_1 and A_2 actually realized are

$$A_1 = 0 - |z| - B_b/\sigma_b$$

$$A_2 = 0 + |z| - B_n/\sigma_b$$

3. DISCUSSION

First we consider bias relative to β. The relative biases for different parameters are shown by three figures attached. Each figure contains three graphs -- the one on the left is for ρ_{XY} = .1, the center one is for ρ_{XY} = .5, and the one on the right is for ρ_{XY} = .9. Each graph shows the relative biases for $\rho_{de} = \pm$ .1, .5, and .9 and for various values of the error variances σ_d^2/σ_X^2 and σ_e^2/σ_Y^2. Figure 1 is for ρ_{Xd} = 0 and ρ_{Xe} = 0, Figure 2 is for ρ_{Xd} = 0.9 and ρ_{Xe} = 0.9, and Figure 3 is for ρ_{Xd} = -0.5 and ρ_{Xe} = -0.5.

The main points of these graphs are: (1) for given ρ_{Xd}, ρ_{Xe}, and ρ_{de}, the greater the ρ_{XY}, the smaller the relative bias for the ranges of the error variances considered. But the magnitude of the relative bias is substantial even when ρ_{XY}=0.9 ρ_{Xd} = 0, and ρ_{Xe} = 0 (see Figure 1). As ρ_{Xd} and ρ_{Xe} increase, the relative biases increase -- with much more variation over different ρ_{de}'s for low ρ_{XY} than for high ρ_{XY}. For $\rho_{Xd}<0$ and $\rho_{Xe}<0$, the relative bias varies more over different ρ_{de}'s for given ρ_{XY} than the relative bias does for $\rho_{Xd}>0$ and $\rho_{Xe}>0$. (2) For given ρ_{Xd} and ρ_{Xe} the relative bias varies more for low values of σ_d^2/σ_X^2 and ρ_{XY} than for higher values of the same parameters; conversely, the relative bias varies more for high values of ρ_{de} and σ_e^2/σ_Y^2 than for lower values of ρ_{de} and σ_e^2/σ_Y^2. When $\rho_{Xd}<0$ and $\sigma_{Xe}<0$ the relative bias is highly sensitive to change in σ_d^2/σ_X^2, σ_e^2/σ_Y^2 and ρ_{de} (see Figure 3).

Next we consider the effect of the bias on interval estimation of β and on hypothesis testing. It is well known that if an estimator is biased the probability of including the parameter in the confidence interval is reduced and that the amount of loss in probability in general is quite serious as the bias relative to the standard error of the estimator is greater than 0.2. Table 1 presents the actual probability realized for a 95 per cent confidence interval for β. The data presented in this table are for $\rho_{XY} = .55$ only. According to the table, the probability decreases as $|\rho_{Xd}|$ increases for a given sample size (n) and for a given ρ_{de}, except for $\rho_{de} = 0$ and $-.56 \leq \rho_{Xd} \leq -.86$. The probability also decreases rather rapidly as n increases (except for $\rho_{de} = 0$ and $\rho_{Xd} = -.86$). This is expected since σ_b^2 gets smaller as $\Sigma(x-\bar{x})^2$ decreases when n increases, whereas the bias B_b is independent of n.

In the case of hypothesis testing, the same consequence is realized as in the case of interval estimation. This time, however, the probability of accepting the null hypothesis when it is indeed true.

In short, what we have shown and reemphasized (on the basis of the model which is more general and perhaps more realistic) is the danger of using the OLSE of β when both variables are subject to errors of measurement. As many statisticians are aware, there has been some progress, however limited, in finding ways of improving estimation methods and a way of actually assessing the various error parameters, but there is much need for more research in this area.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the useful comments given by the referees and Professor Herbert Phillips of Temple University and the computer assistance by Mr. Michael Legault, Ph.D. student, School of Management, Syracuse University.

REFERENCES

[1] Chai, John J. (1971). "Correlated Measurement Errors and Least Squares Estimator of the Regression Coefficient," Journal of the American Statistical Association, 66, 478-483.

[2] _____ and Frankfurter, G. M. (1973). "A Simulation Study of Errors of Measurement," Unpublished manuscript, Working Paper Series, School of Management, Syracuse University.

[3] Cochran, W. G. (1970). "Some Effects of Errors of Measurement on Multiple Correlation," Journal of American Statistical Association, 65, 22-34.

[4] _____(1968). "Errors of Measurement in Statistics," Technometrics, 10, 637-666.

[5] _____(1963). Sampling Techniques, 2nd ed., New York: John Wiley & Sons, Inc., 1963.

[6] DeGracie, J. S. and Fuller, W. A. (1972). "Estimation of the Slope and Analysis of Covariance when the Concommitant Variable is Measured with Error," Journal of the American Statistical Association, 67, 930-937.

[7] Hansen, M. H., Hurwitz, W. N. and Pritzker,L. (1964) "The Estimation and Interpretation of Gross Differences and the Simple Response Variance," Contributions to Statistics Presented to Professor P. S. Mahalanobis on the Occasion of His 70th Birthday, Calcutta: Pergamon Press.

[8] Horwitz, D. G., and Koch, Gary G. (1969). "The Effect of Response Errors on Measures of Association," in N. L. Johnson and H. Smith, Jr., eds, New Development Sampling, New York: John Wiley & Sons, Inc., 247-282.

[9] Johnston, J. (1960). Econometric Methods, first ed., 148-150, New York: McGraw-Hill Book Company.

[10] Koch, Gary G. (1969). "The Effects of Non-Sampling Errors on Measures of Association in 2 x 2 Contingency Tables," Journal of the American Statistical Association, 64, 851-864.

[11] Lindley, D. V. (1947). "Regression Lines and the Linear Functional Relationship," Journal of the Royal Statistical Society, Supplement, 218-244.

TABLE 1 ACTUAL CONFIDENCE LEVEL

REALIZED FOR 95 PER CENT CONFIDENCE

(For $\rho_{XY} = .55$)

		Sample Size			
ρ_{de}	ρ_{Xd}	10	25	50	100
.43	.86	.7985	.4840	.2709	.0505
	.56	.8555	.6406	.3707	.1075
	0	.9305	.8835	.7612	.6141
	-.56	.9333	.8817	.8300	.6985
	-.86	.4920	.0250	0	0
0	.86	.7157	.2946	.0401	0
	.56	.7580	.3783	.0401	.0000
	0	.8395	.5987	.2062	.0708
	-.56	.9219	.8300	.6772	.4840
	-.86	.9498	.9477	.9449	.9441
-.43	.86	.5714	.1492	.0150	0
	.56	.6217	.1922	.0078	0
	0	.6844	.3372	.0559	.0018
	-.56	.7454	.2451	.0582	.0021
	-.86	.7190	.1635	.0154	0

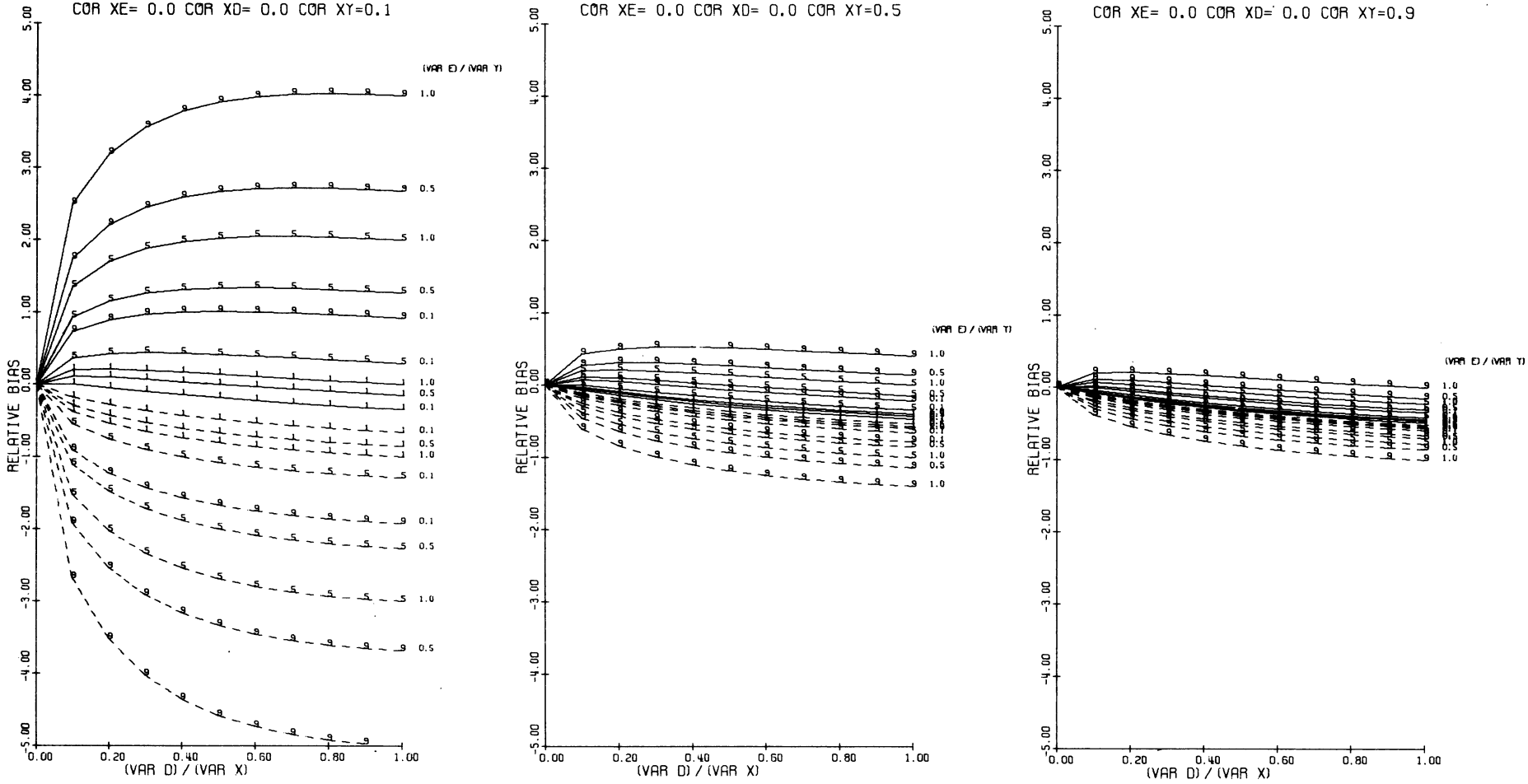

Figure 1 The relative bias of the OLSE for $\rho_{Xe} = 0$, $\rho_{Xd} = 0$, and for selected values of ρ_{de}.

*Remark: Solid lines in Figures represent the relative bias for positive ρ_{de} and the broken lines show the relative bias for negative ρ_{de}. The numbers on the lines indicate the magnitudes of ρ_{de}. For example, 9 means $\rho_{de} = .9$ if it is on a solid line, it means $\rho_{de} = -.9$ if on a broken line. The numbers at the end of curves indicate σ_e^2/σ_Y^2.

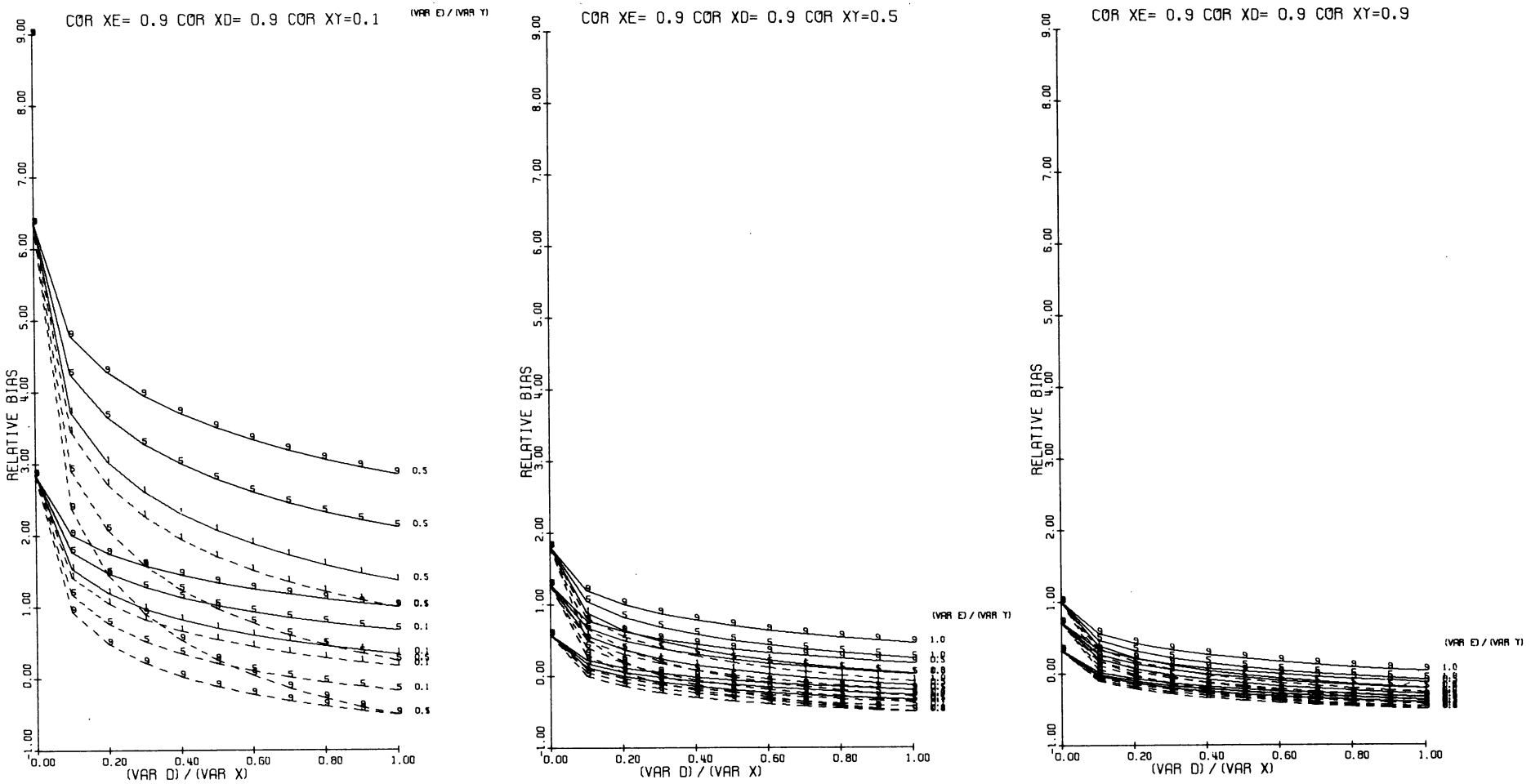

Figure 2 The relative bias of the OLSE for $\rho_{Xe} = .9$, $\rho_{Xd} = .9$, and for selected values of ρ_{de}.

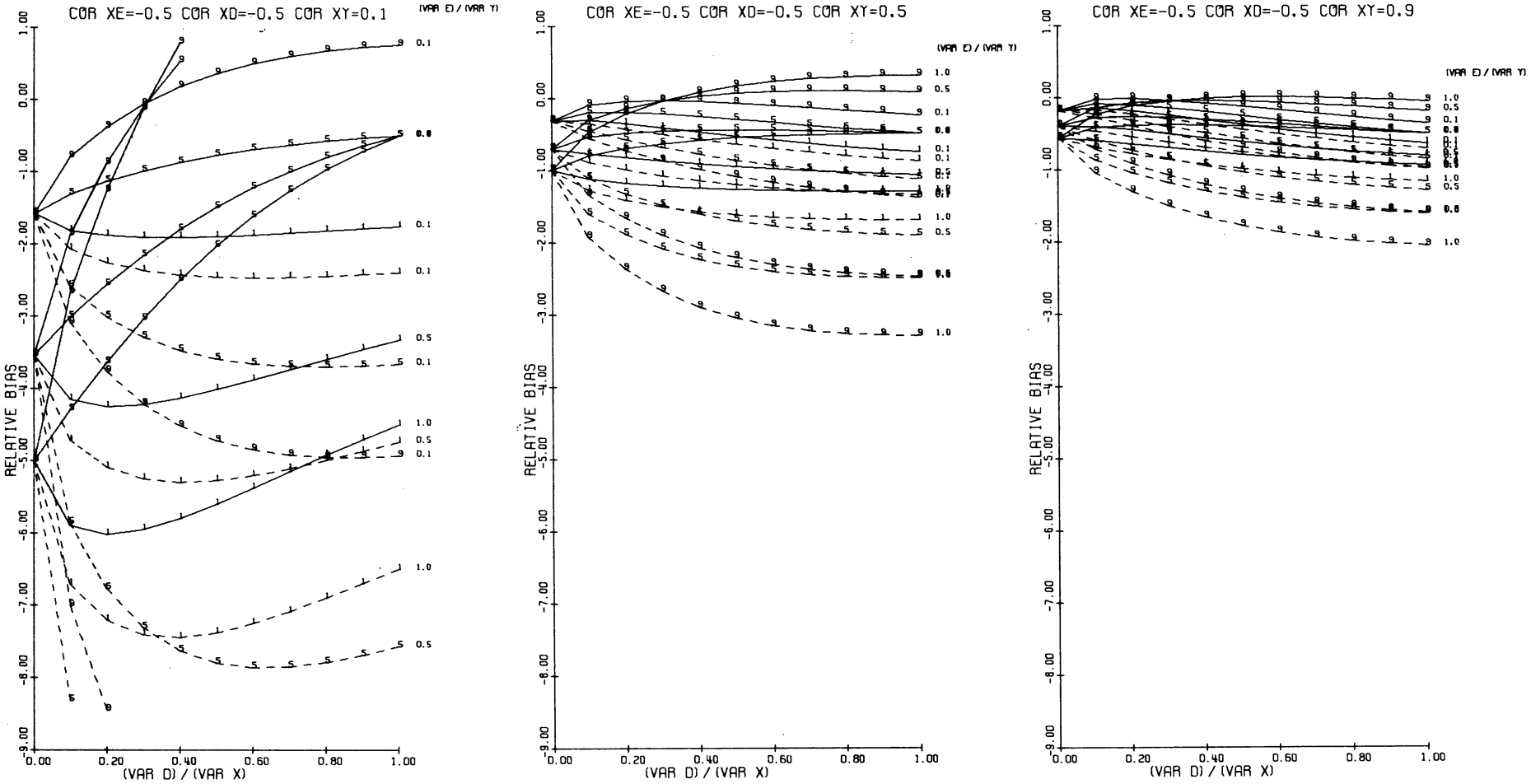

Figure 3 The relative bias of the OLSE for $\rho_{Xe} = -.5$, $\rho_{Xd} = -.5$, and for selected values of ρ_{de}.