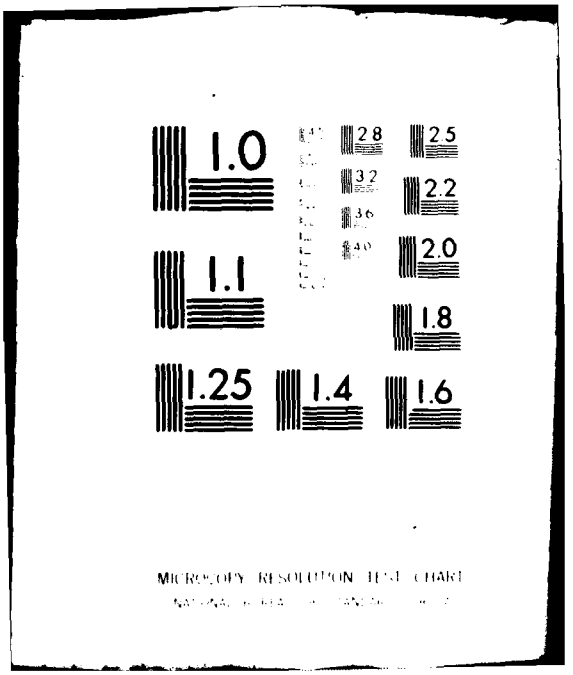
AD-A087 287 CLEMSON UNIV SC DEPT OF MATHEMATICAL SCIENCES F/G 12/1
MEASUREMENT ERROR IN REGRESSION ANALYSIS.(U)
MAY 78 A MITRA, K ALAM N00014-75-C-0451
UNCLASSIFIED N-97 NL

END
DATE FILMED
9-80
DTIC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART

LEVEL II

(11) LEVEL II

MEASUREMENT ERROR IN REGRESSION ANALYSIS

AMITAVA MITRA AND KHURSHEED ALAM

MAY 8, 1978

University of Southern California and Clemson University

Technical Report #281

Report N-97

Research Supported By
THE OFFICE OF NAVAL RESEARCH
Task NR ~~042-271~~ 047-202 Contract N00014-75-C-0451

MEASUREMENT ERROR IN REGRESSION ANALYSIS

Amitava Mitra and Khursheed Alam
University of Southern California and Clemson University

ABSTRACT

Consider the linear regression model $Y = X\theta + \varepsilon$ where Y denotes a vector of n observations on the dependent variable, X is a known matrix, θ is a vector of parameters to be estimated and ε is a random vector of uncorrelated errors. If X'X is nearly singular, that is if the smallest characteristic root of X'X is small then a small perturbation inthe elements of X, such as due to measurement errors, induces considerable variation in the least squares estimate of θ. In this paper we examine for the asymptotic case when n is large the effect of perturbation with regard to the bias and mean squared error of the estimate.

Key words: Linear regression; least squares estimate; mean squared error.

AMS Classification: 62J05

*The authors work was supported by the Office of Naval Research under Contract N00014-75-C-0451.

1. Introduction. Consider the linear regression model

$$Y = X\theta + \varepsilon \tag{1.1}$$

where Y is a n×1 vector of observations, X is a fixed n×p matrix of rank p, θ is a p×1 vector of unknown parameters to be estimated and ε is a n×1 vector of random errors. Let the components of ε be uncorrelated and identically distributed with mean zero and variance σ^2, say. Let $\lambda_1,\ldots,\lambda_p$ denote the characteristic roots of X'X, where prime denotes the transpose of a matrix. The least squares estimate of θ and the sum of mean squared errors (SMSE) of the components of θ are given by

$$\hat{\theta} = (X'X)^{-1}X'Y \tag{1.2}$$

$$\begin{aligned} \text{SMSE } \hat{\theta} &= E(\hat{\theta}-\theta)'\ (\hat{\theta}-\theta) \\ &= \sigma^2 \sum_{i=1}^{p} \lambda_i^{-1}. \end{aligned} \tag{1.3}$$

Clearly, $\hat{\theta}$ is an unbiased estimator of θ. From (1.3) it is seen that if X'X is nearly singular, that is if one or more of the values of λ_i is small then $\hat{\theta}$ is unstable in the sense that the variance of some of the components of $\hat{\theta}$ is large. A small value of λ_i may arise from certain interrelationship between the independent variables of the linear model. The relation is called multicollinearity in econometrics.

Suppose that the elements of X are subjected to small random perturbations, such as due to measurement errors. From (1.2) it is clear that the least square estimator $\hat{\theta}$ is no more unbiased

for θ. Beaton, Rubin and Barone (1976) considered a set of data proposed by Longley (1967) for regression analysis to find the effect of perturbation. They introduced perturbation as round-off errors in the numerical values of the elements of X. From an extensive empirical study they found that the regression analysis could be very sensitive to small perturbations. The authors have concluded from their study that "the computer program is often not the most important factor in computing regression analysis, and that the best thing a program can do for some problems is to refuse to complete the calculations". The conclusion seems to be naive (see Dent and Cavendar (1977) and Espasa (1977) for comments on the authors' paper). The problem arises from the choice of the estimator, namely, the least squares estimator which is unstable when the design matrix X'X is nearly singular. The difficulty can be overcome by choosing some other estimator, such as, the "ridge" estimator, given by $\delta = (X'X+KI)^{-1}X'Y$, where K is a positive number. But then δ is not unbiased.

In this paper we examine the behavior of the least squares estimator when n is large and X is subjected to a random perturbation. Formulas are given for the asymptotic bias and variance. The relation between the bias and the eigen values of X'X is shown through a canonical representation of the parameter θ. It is seen that the smaller the eigen value, the larger is the associated bias. The given formulas are checked with an empirical result obtained by the Monte Carlo method.

In a recent paper, Stewart (1977) has given an upper bound on the deviation of the least squares estimator due to a given perturbation in X. But Stewart's method is not applicable to the derivation of the results given in this paper.

2. Main results. Let F denote the perturbation matrix. That is, X+F represents the perturbed matrix of the independent variables of the linear model (1.1). Suppose that the elements of F are uncorrelated random variables, distributed independent of ε with mean zero and common variance ν, say. The least squares estimator of θ for the perturbed data set is given by

$$\theta^* = ((X+F)'(X+F))^{-1}(X+F)'Y \tag{2.1}$$

Therefore

$$\begin{aligned} E\theta^* &= E((X+F)'(X+F))^{-1}(X+F)'X\theta \\ &= \theta - E((X+F)'(X+F))^{-1}(X+F)'F\theta \end{aligned} \tag{2.2}$$

where the expectation in the second line on the right side of (2.2) is with respect to the distribution of the perturbation errors. Formula (2.2) gives the bias of θ^*.

Let the rows of the matrix X be extended such that the elements of X are uniformly bounded and the characteristic roots of X'X are given by $\lambda_i = n\,\nu_i + O(n^{\frac{1}{2}})$, where $\nu_1,\ldots,\nu_p$ are a fixed set of positive numbers. Let $\alpha = P\theta$ and $\alpha^* = P\theta^*$, where P is an orthogonal matrix diagonalizing X'X. Multiplying both sides of (2.2) by P and equating the ith component of the resulting vector of each side we have after simplification

$$E\,\alpha_i^* = \alpha_i - \left(\frac{\nu}{\nu_i+\nu} + O(n^{-\frac{1}{2}})\right)\alpha_i. \tag{2.3}$$

Similarly, the variance and mean squared error of α_i^* are given by

$$n\,\mathrm{var}\,\alpha_i^* = \frac{\sigma 2}{\nu_i+\nu}\,(1 + O(n^{-\frac{1}{2}})) \tag{2.4}$$

$$(2.5) \qquad n\,E(\alpha_i^*-\alpha_i)^2 = \frac{\sigma^2}{\nu_i+\nu}\,(1 + 0(n^{-\frac{1}{2}})) + \frac{n\nu^2\alpha_i^2}{(\nu_i+\nu)^2}\,(1 + 0(n^{-\frac{1}{2}})).$$

Therefore

$$(2.6) \qquad \begin{aligned} n\ \text{SMSE}\ \theta^* &= n\,E(\theta^*-\theta)'(\theta^*-\theta) \\ &= n\,E(\alpha^*-\alpha)'(\alpha^*-\alpha) \\ &= [\sum_{i=1}^{p} \frac{\sigma^2}{\nu_i+\nu} + \sum_{i=1}^{p} \frac{n\nu^2\alpha_i^2}{(\nu_i+\nu)^2}](1 + 0(n^{-\frac{1}{2}})). \end{aligned}$$

If $\nu = 0$, that is, if there is no perturbation then $E\,\alpha_i^* = \alpha_i$. From (2.3) it is seen that the relative bias of α_i^* is small if ν is small compared to ν_i, as it should be expected. On the other hand, if ν_i is small compared to ν then the relative bias of α_i^* is nearly equal to -1.

From (2.4) it is seen that for $\nu = 0$ we have $\text{var}\,\alpha_i^* = \frac{\sigma^2}{n\nu_i} = \frac{\sigma^2}{\lambda_i}$ which agrees with the result given in (1.3). To see the relation between the effect of perturbation on the variance of α_i^* and the associated eigen value of X'X, we write (2.4) as follows:

$$(2.7) \qquad n\,\nu\,\text{var}\,\alpha_i^* = \frac{\sigma^2\,\nu}{\nu_i+\nu} + 0(n^{-\frac{1}{2}}).$$

From (2.7) it is seen that the perturbation of X has a stabilizing influence on the least square estimate. But the reduction in the variance should be reckoned with the induced bias.

To verify the asymptotic formulas given above, we have carried out the regression analysis under perturbation with a 16×6 matrix X, obtained from the data proposed by Longley (1967). However, the matrix was modified for certain changes in scale and origin. The characteristic roots of the modified matrix are given by $\lambda_i = 16\nu_i$, where

$$\nu_1 = .2188(10)^{-2},\ \nu_2 = .3705(10)^{-1},\ \nu_3 = .2005(10)^{-1}$$

$$\nu_4 = .1118(10)^{2}\ .\ \nu_5 = .1282(10)^{3}\ ,\ \nu_6 = .3596(10)^{4}.$$

From the given values of ν_i we generate as follows an n×p matrix Z for large n such that the characteristic roots of Z'Z are approximately given by $\lambda_i = n\nu_i + 0(n^{\frac{1}{2}})$: Generate a p-component vector U whose components are identically and independently distributed as N(0,1). Compute

$$T = P'\ \sqrt{D}\ U$$

where P is the orthogonal matrix diagonalizing X'X and D denotes the diagonal matrix with diagonal elements equal to ν_i, $i = 1,\ldots,6$. Generate n independent values of T and set them equal to the columns of Z'.

For each Z we generate the error vector ε whose components are independently and identically distributed as N(0,1), that is, $\sigma = 1$. Then we compute Y from the formula $Y = Z\theta + \varepsilon$, where the components of θ are given by

$$\theta_i = .0151,\ \theta_2 = -.3582,\ \theta_3 = -.2020$$

$$\theta_4 = -.1033,\ \theta_5 = -.5110,\ \theta_6 = .1829.$$

The value of θ given above is the least square estimate of θ computed from the data given by Longley. For the discussion of this paper any other value of θ could have been assumed as well.

The matrix Z is perturbed by adding to each element of Z independent values of a random variable ξ, distributed uniformly on $(-\frac{1}{2}, \frac{1}{2})$, giving $\nu = \frac{1}{12}$.

The results of the regression analysis are shown in Table I below. The figures given in the table for the asymptotic bias and mean squared error of the least squares estimate are obtained from

the formulas (2.3) and (2.5). The figures for the empirical values given in the table are each based on 500 simulations. They were found to be fairly accurate, by checking duplicate values. It is seen from the table that there is fair agreement between the theoretical and empirical figures.

Table I - Asymptotic (Asym) and Empirical (Emp) values of $E\,\alpha_i^* - \alpha_i$ and $nMSE(\alpha_i^*)$ for $\nu = \frac{1}{12}$ and $n = 500$.

	$E\alpha_i^* - \alpha_i$		$nMSE(\alpha_i^*)$	
	Asym	Emp	Asym	Emp
i=1	-.3713	-.3914	80.6393	76.7583
2	.3028	.4482	54.1350	50.6663
3	.0027	.0596	.4825	1.9906
4	.0004	.0353	.0889	.7266
5	-.0001	-.0253	.0078	.3884
6	-	-.0025	.0003	.0047

- Denotes insignificant figure

References

[1] Beaton, A.E., Rubin, D.B. and Barone, J.L. (1976). The acceptability of regression solutions: Another look at computational accuracy. Jour. Amer. Statist. Assoc. (71) 158-168.

[2] Dent, W.T. and Cavendar, D.C. (1977). More on computational accuracy in regression. Jour. Amer. Statist. Assoc. (72) 598-602.

[3] Espasa, A. (1977). A note on the acceptability of regression solutions: Another look at computational accuracy. Jour. Amer. Statist. Assoc. (72) 602.

[4] Longley, J.W. (1967). An appraisal of least squares program for the electronic computer from the point of view of the user. Jour. Amer. Statist. Assoc. (62) 819-841.

[5] Stewart, G. W. (1977). On the perturbation of pseudo-inverses, projections and linear least squares problems. Siam Review (19) 634-662.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER (14) N-97 TR-281	2. GOVT ACCESSION NO. AD-A087287
4. TITLE (and Subtitle) (6) Measurment Error in Regression Analysis	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER 281
7. AUTHOR(s) (10) Amitava Mitra ~~and~~ Khursheed Alam	8. CONTRACT OR GRANT NUMBER(s) (15) N00014-75-C-0451
9. PERFORMING ORGANIZATION NAME AND ADDRESS Clemson University Dept. of Mathematical Sciences Clemson, South Carolina 29631 407 183	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 047-202 NR ~~042-271~~
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Code ~~436~~ 434 Arlington, Va. 22217	(11) 12. REPORT DATE MAY [illegible]78 13. NUMBER OF PAGES 8 (12) 12
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Linear regression; least squares estimate; mean squared error.

epsilon theta epsilon

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Consider the linear regression model $Y = X\theta + \varepsilon$ where Y denotes a vector of n observations on the dependent variable, X is a known matrix, θ is a vector of parameters to be estimated and ε is a random vector of uncorrelated errors. If X'X is nearly singular, that is if the smallest characteristic root of X'X is small then a small perturbation in the elements of X, such as due to measurement errors, induces considerable variation in the least squares estimate of θ. In this paper we examine for the asymptotic case when n is large the effect of perturbation with regard to the bias and mean squared error of the estimate.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

407183 JOB