

AD-A072 597

NORTH CAROLINA UNIV AT CHAPEL HILL INST OF STATISTICS

F/G 12/1

AN INVESTIGATION INTO THE EFFECTS OF ASYMMETRY ON ROBUST ESTIMA--ETC(U)

AUG 78 R J CARROLL

AFOSR-75-2796

UNCLASSIFIED

MIMEO SER-1172

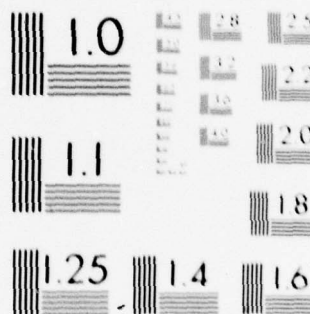
AFOSR-TR-78-1431

NL

| OF |  
AD  
A072597



END  
DATE  
FILMED  
9-79  
DDC



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR-TR. 78-1431

LEVEL II

AD A 072597

DDC FILE COPY

An Investigation into the Effects of  
Asymmetry on Robust Estimates of Regression

Raymond J. Carroll

Institute of Statistics Mimeo Series #1172

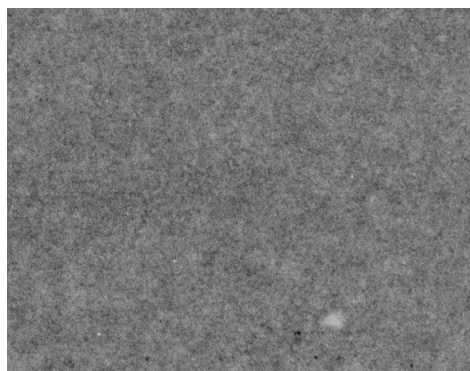
August 1978

DDC  
RECEIVED  
AUG 13 1979  
D

DEPARTMENT OF STATISTICS  
Chapel Hill, North Carolina

DISTRIBUTION STATEMENT A

Approved for public release;  
Distribution Unlimited



| 1. REPORT DOCUMENTATION PAGE                                                                                                                                                                                                              |                  | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM                 |                               |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------|-------------------------------------------------------------|-------------------------------|
| 18                                                                                                                                                                                                                                        | AFOSR-TR-78-1431 | 2. GOVT ACCESSION NO.                                       | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)                                                                                                                                                                                                                   |                  | 5. TYPE OF REPORT & PERIOD COVERED                          |                               |
| 6. AN INVESTIGATION INTO THE EFFECTS OF ASYMMETRY ON ROBUST ESTIMATES OF REGRESSION.                                                                                                                                                      |                  | 9. Interim rept.                                            |                               |
| 7. AUTHOR(s)                                                                                                                                                                                                                              |                  | 8. CONTRACT OR GRANT NUMBER(s)                              |                               |
| 10. Raymond J. Carroll                                                                                                                                                                                                                    |                  | 15. AFOSR-75-2796                                           |                               |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS                                                                                                                                                                                               |                  | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS |                               |
| University of North Carolina<br>Department of Statistics<br>Chapel Hill, North Carolina 27514                                                                                                                                             |                  | 61102F 2304 A5                                              |                               |
| 11. CONTROLLING OFFICE NAME AND ADDRESS                                                                                                                                                                                                   |                  | 12. REPORT DATE                                             |                               |
| Air Force Office of Scientific Research/NM<br>Bolling AFB, Washington, DC 20332                                                                                                                                                           |                  | 11. August 1978                                             |                               |
| 14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)                                                                                                                                                               |                  | 13. NUMBER OF PAGES                                         |                               |
| 17. A5                                                                                                                                                                                                                                    |                  | 24                                                          |                               |
|                                                                                                                                                                                                                                           |                  | 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 abstract entered in Block 20, if different from Report)                                                                                                                                                    |                  |                                                             |                               |
| 14. MIMED SER-1172                                                                                                                                                                                                                        |                  |                                                             |                               |
| 18. SUPPLEMENTARY NOTES                                                                                                                                                                                                                   |                  |                                                             |                               |
| 19. KEY WORDS (Continue on reverse side if necessary and identify by block number)                                                                                                                                                        |                  |                                                             |                               |
| Regression, Robustness, Asymmetry, Variance Estimate, Jackknife, Monte-Carlo M-estimates.                                                                                                                                                 |                  |                                                             |                               |
| 20. ABSTRACT (Continue on reverse side if necessary and identify by block number)                                                                                                                                                         |                  |                                                             |                               |
| We investigate the effects of asymmetry errors on robust regression estimates. Theoretical and Monte-Carlo results show that slopes are essentially unaffected but that the intercept has a bias and its variance is difficult to access. |                  |                                                             |                               |

410 064

DM



|                    |                                     |
|--------------------|-------------------------------------|
| Accession For      |                                     |
| NTIS Grant         | <input checked="" type="checkbox"/> |
| DDC TAB            | <input checked="" type="checkbox"/> |
| Unannounced        | <input type="checkbox"/>            |
| Justification      |                                     |
| By _____           |                                     |
| Distribution/      |                                     |
| Availability Codes |                                     |
| Dist.              | Avail and/or special                |
| A                  |                                     |

# LEVEL

2

## AN INVESTIGATION INTO THE EFFECTS OF ASYMMETRY ON ROBUST ESTIMATES OF REGRESSION

by

Raymond J. Carroll\*  
University of North Carolina at Chapel Hill

### ABSTRACT

We investigate the effects of asymmetry errors on robust regression estimates. Theoretical and Monte-Carlo results show that slopes are essentially unaffected but that the intercept has a bias and its variance is difficult to assess.

DDC  
RECEIVED  
AUG 13 1979  
D

Key Words and Phrases: Regression, Robustness, Asymmetry, Variance estimation, jackknife, Monte-Carlo M-estimates.

\* Research supported by the Air Force Office of Scientific Research under Grant AFOSR-75-2796.

### DISTRIBUTION STATEMENT A

Approved for public release;  
Distribution Unlimited

## Introduction

The theoretical results and Monte-Carlo studies in the area of robustness have in the main focused on symmetric distributions (Andrews, et al (1972)) or procedures which are not scale invariance (which effectively eliminates most problems due to asymmetry when the number of dimensions in the problem is fixed). Recently, Huber (1973) and Bickel (1978) have examined situations in which the asymmetry of errors can lead to quite complicated results. In this paper we study the effects of asymmetric errors in regression.

A major difficulty with considering asymmetric errors has been that location (intercept) is not uniquely defined. However, asymmetric data do occur and there are situations where data transformation to achieve symmetry either make no sense or are not possible. In regression, it might be conjectured that asymmetry has different effects on intercept and slope (see Section 3); if so, there will be situations where one might invest much effort in data transformations, when the parameters of interest are not influenced by the asymmetry.

Carroll (1978c) considered asymmetric errors in regression and illustrated his results by means of a Monte-Carlo study, using simple linear regression with a uniform design. These results indicate that asymmetry effects robust M-estimates of regression only through the intercept term, which may be biased and have a variance that cannot be consistently estimated by the usual variance estimates. The purpose of this report is to expand Carroll's (1978c) Monte-Carlo study to a wide variety of designs. The results are almost staggering in their consistency (and confirm the results of Carroll (1978c)), especially in view of the fact that the designs we consider range from balanced to unbalanced with a large amount of multicollinearity.

In Section 2, we review the theory of M-estimates as it applies to situations where the errors are possibly asymmetric. In Section 3, we report the Monte-Carlo results, while in Section 4 we present our conclusions.

### M-estimates

In the one sample problem, we have a sample  $X_1, X_2, \dots, X_n$  from a distribution function  $F$ . Our aim is to estimate the center of the distribution.

Huber (1964) defines the center  $\theta$  by

$$(2.1) \quad \int \psi(x - \theta) dF(x) = 0,$$

where  $\psi$  is a skew-symmetric ( $\psi(x) = -\psi(-x)$ ) nondecreasing function. If  $\psi(x) = x$ ,  $\theta$  is the population mean. Huber (1964) then proposed to estimate  $\theta$  by solving

$$(2.2) \quad \int \psi(x - \theta) dF_n(x) = n^{-1} \sum_{i=1}^n \psi(X_i - \theta) = 0,$$

with the solution denoted by  $T_n$ . If  $\psi(x) = x$ , we obtain the sample mean, which is well-known to lack robustness against outliers. One possible choice of  $\psi$  to achieve this robustness is

$$\psi(x) = \max(-k, \min(x, k)),$$

where in our Monte-Carlo study we take  $k = 2$ . The estimate obtained by solving equation (2.2) is not scale equivariant; to achieve this property, Huber (1977) proposes solving the system of equations



$$(2.3) \quad n^{-1} \sum_{i=1}^n \psi((X_i - T_n)/s_n) = 0$$

$$(2.4) \quad (n-1)^{-1} \sum_{i=1}^n \psi^2((X_i - T_n)/s_n) = a = E_{\Phi} \psi^2(X_1),$$

where the last expectation is taken under the standard normal distribution. If  $F$  is symmetric, it can be shown that if  $T_n \xrightarrow{P} T(F)$ ,  $s_n \xrightarrow{P} \sigma(F)$ , then

$$n^{1/2}(T_n - T(F)) \xrightarrow{L} N(0, A(\psi, F)),$$

where the asymptotic variance is  $(T(F) = 0, \sigma(F) = 1)$

$$A(\psi, F) = \int \psi^2(x) dF(x) / \{ \int \psi'(x) dF(x) \}^2.$$

This suggests that the variance of  $T_n$  be estimated by

$$(2.5) \quad D_n = s_n^2 (n-1)^{-1} \sum_{i=1}^n \psi^2((X_i - T_n)/s_n) / \{ n^{-1} \sum_{i=1}^n \psi^1((X_i - T_n)/s_n) \}^2,$$

as is suggested by Gross (1976).

When  $F$  is asymmetric, these results are not true and in fact Carroll (1978a) (extended to multivariate situations in Carroll (1978b)) shows the following result.

Lemma 1. Suppose that for some constants  $T(F)$ ,  $\sigma(F)$  we have that  $T_n \xrightarrow{P} T(F)$ ,  $n^{1/2}(s_n - \sigma(F)) = O_p(1)$  and  $E\psi((X_1 - T(F))/\sigma(F)) = 0$ . Taking  $T(F) = 0$ ,  $\sigma(F) = 1$  (without loss of generality), for  $\psi$  smooth,

$$(2.6) \quad (E_F \psi'(X_1))T_n = n^{-1} \sum_{i=1}^n \psi(X_i) + (1-s_n)E_F X_1 \psi'(X_1) + O_p(n^{-1}).$$

□

This result says that  $D_n$  will not be a consistent estimate of the variance of  $n^{1/2}T_n$  if  $F$  is asymmetric. Carroll (1978c) shows by examples that there exist situations for which

$$\frac{E D_n}{\text{Var}(n^{1/2}T_n)} < .65 ,$$

and there are presumably distributions where this is worse. The question we want to answer is how asymmetry will influence robust regression estimates.

The model we consider is (the use of  $\sigma_0$  will become clear later)

$$(2.7) \quad y_i = \underline{x}_i \underline{\beta}_0 + \varepsilon_i \sigma_0 \quad (i = 1, \dots, n) ,$$

where the  $\varepsilon_i$  are i.i.d. random variables with  $E\psi(\varepsilon_1) = 0$  and  $\underline{x}_i = (1 \ x_{i1} \dots x_{ip})$ . We consider a version of Huber's Proposal 2 (Huber (1977), p. 37), which involves solving the equations

$$(2.8) \quad n^{-1} \sum_{i=1}^n \psi((y_i - \underline{x}_i \underline{\beta}_n)/s_n) = 0$$

$$(2.9) \quad (n-p)^{-1} \sum_{i=1}^n \psi^2((y_i - \underline{x}_i \underline{\beta}_n)/s_n) = a .$$

While we will assume the  $\underline{x}_i$  are constants, the first two conditions of Lemma 2 (to follow) are reasonable and may be justified by quoting results of Maronna and Yohai (1978). In a subset of their paper, they assume  $(y_1, \underline{x}_1), (y_2, \underline{x}_2), \dots$  is a sample from a distribution function  $P$  with  $\underline{\beta}_0, \sigma_0$  solving

$$E\psi((y - \underline{x} \underline{\beta}_0)/\sigma_0) = 0$$

$$E\psi^2((y - \underline{x} \underline{\beta}_0)/\sigma_0) = a .$$

Defining  $\epsilon_i = (y_i - \underline{x}_i' \beta_0)/\sigma_0$ , they show essentially that if the  $\epsilon_i$  are independent of  $\underline{x}_i$  and if  $\beta_n, s_n$  satisfy (3.2) and (3.3) (the latter with  $(n-p)$  replaced by  $n$ ), then  $n^{1/2}(\beta_n - \beta_0)$  and  $n^{1/2}(s_n - \sigma_0)$  are asymptotically normally distributed.

The proof of our result involves Taylor expansions along the lines of Carroll (1978a, 1978b) and is omitted. Recall that our  $\underline{x}_i$  are non-stochastic.

Lemma 2. Suppose that

$$n^{1/2}(\beta_n - \beta_0) = O_p(1)$$

$$n^{1/2}(s_n - \sigma_0) = O_p(1)$$

$$n^{-1}X'X \rightarrow V(\text{positive definite})$$

$$n^{-1} \sum_{i=1}^n \underline{x}_i \rightarrow (1, 0, 0, \dots, 0) = \underline{w}$$

Then, for  $\psi$  sufficiently smooth,

$$\begin{aligned} (2.10) \quad & (a_3 V - (a_1 a_4 / a_2) \underline{w}' \underline{w}) (\beta_n - \beta_0) / \sigma_0 \\ & = n^{-1} \sum_{i=1}^n \{ \underline{x}_i' \psi(\epsilon_i) - (a_1 / a_2) \underline{w}' (\psi^2(\epsilon_i) - a) \} \\ & + O_p(n^{-1}), \end{aligned}$$

where

$$a_1 = E \epsilon_1 \psi'(\epsilon_1)$$

$$a_3 = E \psi'(\epsilon_1)$$

$$a_2 = 2E \epsilon_1 \psi(\epsilon_1) \psi'(\epsilon_1)$$

$$a_4 = 2E \psi(\epsilon_1) \psi'(\epsilon_1) .$$

In simple linear regression with

$$V = \begin{bmatrix} 1 & 0 \\ 0 & v_1 \end{bmatrix}$$

we obtain that if  $\underline{\beta}'_0 = (\beta_{\text{int}}, \beta_{\text{slope}})$  and  $\underline{\beta}'_n = (\hat{\beta}_{\text{int}}, \hat{\beta}_{\text{slope}})$ , then

Corollary. Under the conditions of Lemma 3, for simple linear regression

$$(2.11) \quad (\hat{\beta}_{\text{int}} - \beta_{\text{int}})/\sigma_0 = (a_3 - (a_1 a_4/a_2))^{-1} n^{-1} \sum_{i=1}^n \{\psi(\varepsilon_i) - (a_1/a_2)(\psi^2(\varepsilon_i) - a)\} + o_p$$

$$(2.12) \quad (\hat{\beta}_{\text{slope}} - \beta_{\text{slope}})/\sigma_0 = (a_3 v_1)^{-1} n^{-1} \sum_{i=1}^n x_{i1} \psi(\varepsilon_i) + o_p(n^{-1}). \quad \square$$

Similar results hold for the general regression problem.

Lemma 2 says that, at least theoretically for large  $n$  when the dimension of the problem remains fixed, the effect of asymmetry of errors on robust estimators of regression occurs mainly in the intercept. If  $F$  is symmetric, Lemma 2 says that since  $a_1 = 0$ ,

$$\begin{aligned} & \text{Var}(n^{1/2}(\beta_n - \beta)/\sigma_0) \\ & \approx \frac{E_F \psi^2(X_1)}{\{E_F \psi^1(X_1)\}^2} V^{-1}, \end{aligned}$$

and it has been suggested (Gross (1977)) that we estimate this variance by



$$(2.13) \quad D = \frac{\frac{s_n^2}{n-p} \sum_{i=1}^n \psi^2(r_i)}{\left\{ \frac{1}{n} \sum_{i=1}^n \psi^1(r_i) \right\}^2} (n^{-1} X'X)^{-1},$$

where  $r_i = (y_i - \underline{x}_i \beta_n) / s_n$ . What is clear from Lemma 2 and the Corollary are the following.

#### Theoretical Conclusions

(i) The effect of asymmetry of errors on robust regression estimates is evidenced only in the intercept, which will tend to be "biased" and have a variance which is larger than expected from the symmetric case.

(ii)  $D_n$  will be a consistent estimate of the variance of the slopes, but will be generally inconsistent for the variance of the intercept.

(iii)  $D_n$  will be a consistent estimate of the covariance.

In the next section we illustrate the results with a large Monte-Carlo study.

#### Monte-Carlo Experiment

Let  $Z$  be a standard normal random variable. The five distributions presented here are as follows:

| <u>Type</u>                     | <u>Code</u> |
|---------------------------------|-------------|
| $Z$                             | $Z$         |
| $Z + .10Z^2$                    | $.10Z^2$    |
| $Z + .50Z^2$                    | $.50Z^2$    |
| Negative Exponential, mean 1.25 | NE          |
| $.50 \text{ Exp}(Z)$            | EXP(Z)      |

The second (.10Z<sup>2</sup>) is only slightly skewed and was chosen to be reasonably representative of the class of distributions close to, but not exactly, symmetric. Such data might arise for example from data transformations which only to achieve approximate symmetry. If  $N$  is the number of Monte-Carlo experiments and  $Y_1, \dots, Y_N$  the realized value of an estimate, the Monte-Carlo variance is defined to be

$$(3.1) \quad \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y}_N)^2 .$$

All the models are of the form

$$\underline{Y} = \underline{X} \underline{\beta} + \underline{\epsilon},$$

where  $\underline{\beta}' = (\beta_0, \beta_1, \dots, \beta_p)$ . Let  $\hat{\beta}_1, \dots, \hat{\beta}_N$  denote the  $N$  realized values of the robust estimate of  $\underline{\beta}$ , with

$$\hat{\beta}_i' = (\hat{\beta}_{i0}, \dots, \hat{\beta}_{ip}).$$

Form the diagonal matrix  $W$  with  $j$ th diagonal element

$$\frac{1}{N} \sum_{i=1}^N \hat{\beta}_{ij}^2 - \left( \frac{1}{N} \sum_{i=1}^N \hat{\beta}_{ij} \right)^2 .$$

Then, in all our tables,  $E(\beta_j)$  denotes the average value of the  $\hat{\beta}_{ij}$ , i.e.,

$$E \beta_j = \frac{1}{N} \sum_{i=1}^N \hat{\beta}_{ij} .$$

The term  $V(\beta_j)$  is the standardized Monte-Carlo variance of the robust estimate of  $\beta_j$  and is the  $i$ th diagonal element of the matrix

$$X'XW.$$

Under symmetry,  $V(\beta_j) \approx E \psi^2(\epsilon_1) / \{E \psi^1(\epsilon_1)\}^2$ , so that the usual estimate of  $V(\beta_j)$  (Gross (1977)) is denoted by  $\hat{V}(\beta_j)$  and is defined by

$$\hat{V}(\beta_j) = \frac{\frac{s_n^2}{n-p} \sum_{i=1}^n \psi^2(r_i)}{\frac{1}{n} \sum_{i=1}^n \psi^1(r_i)^2}.$$

Hence, the tables contain the following information (NITER = number of iterations).

(i)  $E \beta_j$  = Average value of the robust estimate obtained in the study.

(ii)  $V(\beta_j)$  = The Monte-Carlo variance of the estimate of  $\beta_j$ .

To obtain the "true" Monte-Carlo variance of  $\beta_j$ , merely multiply  $V(\beta_j)$  by the  $j^{\text{th}}$  diagonal element of  $(X'X)^{-1}/N$ .

(iii)  $\hat{V}(\beta_j)$  = The usual estimate of  $V(\beta_j)$  obtained assuming symmetry.

(iv) Ratio =  $\hat{V}(\beta_j)/V(\beta_j)$  = Ratio of estimated variance to true variance.

The following designs were considered in our Monte-Carlo experiment. In all cases, the number of iterations is 1200.

Design #1. Here we have

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \epsilon_{ijk}, \text{ where}$$

$$i = 1, 2$$

$$j = 1, 2, 3$$

$$k = 1, 2, 3, 4 \quad (\text{Hence } n = 24)$$

$$\alpha_1 + \alpha_2 = 0$$

$$\gamma_1 + \gamma_2 + \gamma_3 = 0.$$

This is a balanced  $2 \times 3$  design. In Table 1 we call  $\beta_0 = \mu$ ,  $\beta_1 = \alpha_1$ ,  $\beta_2 = \gamma_1$ ,  $\beta_3 = \gamma_2$ . The true values used are  $\beta_0 = 2.50$ ,  $\beta_1 = -.50$ ,  $\beta_2 = -1.00$ ,  $\beta_3 = 0.00$ .

Design #2. Here we have simple linear regression

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (i = 1, \dots, 20)$$

with  $\beta_0 = 1.00$ ,  $\beta_1 = .50$  and the values of the  $x_i$  being  $-.95, -.90, \dots, .90, .95$ .

Design #3. This is the same as Design #2 except the  $x$ 's are unbalanced and given by

| $x$    |
|--------|
| -0.349 |
| -0.344 |
| -0.333 |
| -0.318 |
| -0.297 |
| -0.270 |
| -0.239 |
| -0.202 |
| -0.160 |
| -0.113 |
| -0.060 |
| -0.002 |
| 0.060  |
| 0.128  |
| 0.202  |
| 0.281  |
| 0.365  |
| 0.454  |
| 0.549  |
| 0.649  |



Design #4. This is again simple linear regression, but with 40 design points given by

| $\underline{X}$ | $\underline{X}$ |
|-----------------|-----------------|
| -.342           | -.072           |
| -.340           | -.046           |
| -.338           | -.017           |
| -.334           | .012            |
| -.329           | .043            |
| -.322           | .075            |
| -.315           | .108            |
| -.306           | .143            |
| -.295           | .179            |
| -.281           | .216            |
| -.271           | .254            |
| -.257           | .294            |
| -.242           | .335            |
| -.225           | .378            |
| -.207           | .421            |
| -.188           | .466            |
| -.167           | .512            |
| -.145           | .560            |
| -.122           | .606            |
| -.098           | .658            |

Note the highly unbalanced nature of this design.

Design #5. This is a regression

$$Y_i = 1.00 + .50X_{i1} + .25X_{i2} + \epsilon_i,$$

where the  $X_{i2}$  are essentially translates of  $X_{i1}^2$ . There are  $n=20$  design points arrayed in a uniform manner with  $X'X$  being a diagonal matrix.

| $X_{i1}$ | $X_{i2}$ |
|----------|----------|
| -0.34435 | -0.21371 |
| -0.34205 | -0.21370 |
| -0.33745 | -0.21366 |
| -0.33056 | -0.21351 |
| -0.32136 | -0.21317 |
| -0.30987 | -0.21250 |
| -0.29608 | -0.21135 |
| -0.28000 | -0.20953 |
| -0.26161 | -0.20682 |
| -0.24093 | -0.20296 |
| -0.21795 | -0.19766 |
| -0.19266 | -0.19062 |
| -0.16509 | -0.18147 |
| -0.13521 | -0.16985 |
| -0.10304 | -0.15534 |
| -0.06856 | -0.13750 |
| -0.03179 | -0.11584 |
| 0.00728  | -0.08987 |
| 0.04865  | -0.05904 |
| 0.09231  | -0.02279 |
| 0.13827  | 0.01949  |
| 0.18654  | 0.06843  |
| 0.23710  | 0.12470  |
| 0.28995  | 0.18899  |
| 0.34511  | 0.26204  |
| 0.40257  | 0.34460  |
| 0.46232  | 0.43746  |
| 0.52437  | 0.54146  |
| 0.58872  | 0.65744  |
| 0.65537  | 0.78629  |

Design #6. This is the same model as in Design #5 but with a highly unbalanced design.

| $X_{i1}$ | $X_{i2}$ |
|----------|----------|
| -0.96667 | 0.60148  |
| -0.90000 | 0.47704  |
| -0.83333 | 0.36148  |
| -0.76667 | 0.25481  |
| -0.70000 | 0.15704  |
| -0.63333 | 0.06815  |
| -0.56667 | -0.01185 |
| -0.50000 | -0.08296 |
| -0.43333 | -0.14519 |
| -0.36667 | -0.19852 |
| -0.30000 | -0.24296 |
| -0.23333 | -0.27852 |
| -0.16667 | -0.30519 |
| -0.10000 | -0.32296 |
| -0.03333 | -0.33185 |
| 0.03333  | -0.33185 |
| 0.10000  | -0.32296 |
| 0.16667  | -0.30519 |
| 0.23333  | -0.27852 |
| 0.30000  | -0.24296 |
| 0.36667  | -0.19852 |
| 0.43333  | -0.14519 |
| 0.50000  | -0.08296 |
| 0.56667  | -0.01185 |
| 0.63333  | 0.06815  |
| 0.70000  | 0.15704  |
| 0.76667  | 0.25481  |
| 0.83333  | 0.36148  |
| 0.90000  | 0.47704  |
| 0.96667  | 0.60148  |

Note that

$$X'X = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 2.84 & 2.52 \\ 0 & 2.52 & 2.43 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} .033 & 0 & 0 \\ 0 & 4.32 & -4.48 \\ 0 & -4.48 & 5.06 \end{pmatrix}$$

$$(X'X)^{-1} \text{ in correlation form} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -.96 \\ 0 & -.96 & 1 \end{pmatrix} .$$

Hence this design is highly unbalanced with a great deal of multicollinearity (i.e.,  $X_{i1}$  and  $X_{i2}$  are highly correlated).



Design #1

The true values are  $\beta_0 = 2.50$ ,  $\beta_1 = -.50$ ,  $\beta_2 = -1.00$ ,  $\beta_3 = 0.00$ .

|                    | <u>Z</u> | <u>.10Z<sup>2</sup></u> | <u>.50Z<sup>2</sup></u> | <u>NE</u> | <u>EXP(Z)</u> |
|--------------------|----------|-------------------------|-------------------------|-----------|---------------|
| E $\beta_0$        | 2.50     | 2.49                    | 2.44                    | 2.45      | 2.41          |
| V( $\beta_0$ )     | .99      | .98                     | 1.46                    | 1.49      | .71           |
| $\hat{V}(\beta_0)$ | .99      | .99                     | 1.17                    | 1.27      | .52           |
| Ratio              | 1.00     | 1.01                    | .80                     | .85       | .72           |
| E $\beta_1$        | - .50    | - .50                   | - .50                   | - .51     | - .50         |
| V( $\beta_1$ )     | .98      | .99                     | 1.16                    | 1.38      | .51           |
| $\hat{V}(\beta_1)$ | .99      | .99                     | 1.17                    | 1.27      | .52           |
| Ratio              | 1.01     | 1.00                    | 1.01                    | .93       | 1.01          |
| E $\beta_2$        | -1.00    | -1.00                   | -1.00                   | -1.00     | -1.00         |
| V( $\beta_2$ )     | .96      | .97                     | 1.12                    | 1.33      | .50           |
| $\hat{V}(\beta_2)$ | .99      | .99                     | 1.17                    | 1.27      | .52           |
| Ratio              | 1.03     | 1.02                    | 1.05                    | .96       | 1.03          |
| E $\beta_3$        | .01      | .00                     | .01                     | -.02      | .00           |
| V( $\beta_3$ )     | 1.04     | 1.03                    | 1.18                    | 1.30      | .53           |
| $\hat{V}(\beta_3)$ | .99      | .99                     | 1.17                    | 1.27      | .52           |
| Ratio              | .95      | .96                     | .99                     | .98       | .97           |

## Design #2

The true values are  $\beta_0 = 1.00$ ,  $\beta_1 = .50$ .

|                    | $Z$  | $\frac{.10Z^2}{.99}$ | $\frac{.50Z^2}{.92}$ | $\frac{NE}{.94}$ | $\frac{EXP(Z)}{.90}$ |
|--------------------|------|----------------------|----------------------|------------------|----------------------|
| $E \beta_0$        | 1.00 |                      |                      |                  |                      |
| $V(\beta_0)$       | 1.05 | 1.06                 | 1.49                 | 1.50             | .70                  |
| $\hat{V}(\beta_0)$ | 1.00 | .99                  | 1.11                 | 1.24             | .48                  |
| Ratio              | .95  | .93                  | .75                  | .83              | .69                  |
| $E \beta_1$        | .49  | .49                  | .49                  | .48              | .50                  |
| $V(\beta_1)$       | 1.02 | 1.01                 | 1.16                 | 1.25             | .52                  |
| $\hat{V}(\beta_1)$ | 1.00 | .99                  | 1.11                 | 1.24             | .48                  |
| Ratio              | .98  | .98                  | .96                  | .99              | .93                  |

Design #3

The true values are  $\beta_0 = 1.00$ ,  $\beta_1 = .50$ .

|                    | <u>Z</u> | <u>.10Z<sup>2</sup></u> | <u>.50Z<sup>2</sup></u> | <u>NE</u> | <u>EXP(Z)</u> |
|--------------------|----------|-------------------------|-------------------------|-----------|---------------|
| E $\beta_0$        | 1.00     | .99                     | .92                     | .94       | .90           |
| V( $\beta_0$ )     | 1.05     | 1.06                    | 1.49                    | 1.50      | .70           |
| $\hat{V}(\beta_0)$ | 1.00     | .99                     | 1.11                    | 1.24      | .48           |
| Ratio              | .95      | .94                     | .75                     | .83       | .69           |
|                    |          |                         |                         |           |               |
| E $\beta_1$        | .49      | .49                     | .50                     | .48       | .50           |
| V( $\beta_1$ )     | 1.03     | 1.02                    | 1.17                    | 1.24      | .53           |
| $\hat{V}(\beta_1)$ | 1.00     | .99                     | 1.11                    | 1.24      | .48           |
| Ratio              | .97      | .98                     | .95                     | 1.00      | .92           |

Design #4

The true values are  $\beta_0 = 1.00$ ,  $\beta_1 = .50$ .

|                    | <u>Z</u> | <u>.10Z<sup>2</sup></u> | <u>.50Z<sup>2</sup></u> | <u>NE</u> | <u>EXP(Z)</u> |
|--------------------|----------|-------------------------|-------------------------|-----------|---------------|
| E $\beta_0$        | 1.00     |                         | .91                     |           | .88           |
| V( $\beta_0$ )     | 1.02     |                         | 1.46                    |           | .64           |
| $\hat{V}(\beta_0)$ | 1.00     |                         | 1.05                    |           | .43           |
| Ratio              | .98      |                         | .72                     |           | .67           |
| E $\beta_1$        | .50      |                         | .51                     |           | .51           |
| V( $\beta_1$ )     | 1.00     |                         | 1.06                    |           | .43           |
| $\hat{V}(\beta_1)$ | 1.00     |                         | 1.06                    |           | .43           |
| Ratio              | 1.00     |                         | 1.00                    |           | 1.00          |



Design #5

The true values are  $\beta_0 = 1.00$ ,  $\beta_1 = .50$ ,  $\beta_2 = .25$ .

|                    | <u>Z</u> | <u>.10Z<sup>2</sup></u> | <u>.50Z<sup>2</sup></u> | <u>NE</u> | <u>EXP(Z)</u> |
|--------------------|----------|-------------------------|-------------------------|-----------|---------------|
| E $\beta_0$        | 1.00     | .98                     | .91                     | .92       | .89           |
| V( $\beta_0$ )     | .99      | 1.01                    | 1.47                    | 1.51      | .68           |
| $\hat{V}(\beta_0)$ | 1.00     | 1.00                    | 1.02                    | 1.19      | .46           |
| Ratio              | 1.02     | .99                     | .74                     | .79       | .67           |
|                    |          |                         |                         |           |               |
| E $\beta_1$        | .49      | .49                     | .50                     | .50       | .50           |
| V( $\beta_1$ )     | 1.02     | 1.02                    | 1.14                    | 1.24      | .49           |
| $\hat{V}(\beta_1)$ | 1.00     | 1.00                    | 1.09                    | 1.19      | .46           |
| Ratio              | .99      | .98                     | .96                     | .95       | .93           |
|                    |          |                         |                         |           |               |
| E $\beta_2$        | .24      | .25                     | .27                     | .27       | .27           |
| V( $\beta_2$ )     | 1.00     | 1.00                    | 1.12                    | 1.16      | .47           |
| $\hat{V}(\beta_2)$ | 1.00     | 1.00                    | 1.09                    | 1.19      | .46           |
| Ratio              | 1.00     | 1.00                    | .97                     | 1.02      | .98           |

$$X'X = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 9.99 & 0 \\ 0 & 0 & 2.65 \end{pmatrix}$$

$$\text{Corr. Matrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} .03 & 0 & 0 \\ 0 & .10 & 0 \\ 0 & 0 & .38 \end{pmatrix}$$

Design #6

The true values are  $\beta_0 = 1.00$ ,  $\beta_1 = .50$ ,  $\beta_2 = .25$ .

|                    | <u>Z</u> | <u>.10Z<sup>2</sup></u> | <u>.50Z<sup>2</sup></u> | <u>NE</u> | <u>EXP(Z)</u> |
|--------------------|----------|-------------------------|-------------------------|-----------|---------------|
| E $\beta_0$        | 1.00     | .98                     | .91                     | .92       | .89           |
| V( $\beta_0$ )     | .99      | 1.01                    | 1.47                    | 1.51      | .67           |
| $\hat{V}(\beta_0)$ | 1.00     | 1.00                    | 1.09                    | 1.18      | .46           |
| Ratio              | 1.02     | .99                     | .74                     | .79       | .68           |
|                    |          |                         |                         |           |               |
| E $\beta_1$        | .49      | .49                     | .46                     | .48       | .46           |
| V( $\beta_1$ )     | .98      | .97                     | 1.08                    | 1.16      | .45           |
| $\hat{V}(\beta_1)$ | 1.00     | 1.00                    | 1.09                    | 1.18      | .46           |
| Ratio              | 1.02     | 1.02                    | 1.09                    | 1.03      | 1.02          |
|                    |          |                         |                         |           |               |
| E $\beta_2$        | .24      | .26                     | .31                     | .30       | .30           |
| V( $\beta_2$ )     | .97      | .97                     | 1.10                    | 1.17      | .46           |
| $\hat{V}(\beta_2)$ | 1.00     | 1.00                    | 1.09                    | 1.18      | .46           |
| Ratio              | 1.03     | 1.02                    | .99                     | 1.01      | 1.00          |

### Conclusions

The Monte-Carlo results are surprisingly consistent. It appears that, in robust regression, one can accurately estimate slopes and their variances, even if the design is highly unbalanced with considerable multicollinearity; the conclusion holds over a wide class of distributions varying from the normal to a heavy tailed, very skewed distribution ( $\text{EXP}(Z)$ ). However, estimating intercept (and especially its variance) is complex if the distributions are heavily skewed. As in Carroll (1978c), we recommend that if one has an unbalanced design, a heavily skewed error distribution, and wishes to estimate terms involving the intercept, variance could be assessed by using the weighted jackknife of Hinkley (1977).

## REFERENCES

- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton University Press.
- Bickel, P.J. (1978). Using residuals robustly I: tests for heteroscedosticity, nonlinearity. *Ann. Statist.* 6, 266-291.
- Carroll, R.J. (1978a). On almost sure expansions for M-estimates. *Ann. Statist.* 6, 314-318.
- Carroll, R.J. (1978b). On almost sure expansions for multivariate M-estimates. To appear in *J. Mult. Analysis*.
- Carroll, R.J. (1978c). On estimating variances of robust estimators when the errors are asymmetric. Mimeo Series #1172, Department of Statistics, University of North Carolina at Chapel Hill.
- Gross, A.M. (1976). Confidence interval robustness with long-tailed symmetric distributions. *J. Am. Statist. Assoc.* 71, 409-416.
- Gross, A.M. (1977). Confidence intervals for bisquare regression estimates. *J. Am. Statist. Assoc.* 72, 341-354.
- Hinkley, D.V. (1976). On jackknifing in unbalanced situations. Technical Report No. 22, Division of Biostatistics, Stanford University.
- Huber, P.J. (1973). Robust regressions: asymptotics, conjectures, and Monte-Carlo. *Ann. Statist.* 1, 799-821.
- Huber, P.J. (1977). *Robust Statistical Procedures*. SIAM, Philadelphia.
- Jaeckel, L.A. (1972). The infinitesimal jackknife. *Bell Laboratories Memorandum* MM-72-1215-11.
- Maronna, R.A. and Yohai, V.J. (1978). Robust M-estimators for regression with contaminated independent variables. Unpublished manuscript.



