# ROBUST REGRESSION AND PROC JACKREGR

Gerry Hobbs, West Virginia University
Daniel M. Chilko, West Virginia University

## INTRODUCTION

One the most common of all statistical techniques and certainly the most common of the predictive techniques is linear regression. The least squares regression coefficients are computed in many SAS procedures. Some of them are the procedures named NLIN, REG, RSQUARE, RSREG, STEPWISE, and GLM. The estimates are also contained in some of the SAS-ETS software.

It has long been recognized that the estimation of regression coefficients by the method of least squares can lead to results which are not altogether satisfactory. In simple (one independent variable) linear regression situations the problems generally arise when one or more outliers exist in the data set. The nature of least squares calculations are such that extraordinary observations exert an undue influence on values of the estimates. In multiple linear regression problems outliers are also a problem and of course it is harder to identify outliers when several independent variables are present than it is when there is only one. In addition, multiple regression estimates are often unstable even in the absence of outliers due to a condition called collinearity. The crux of the problem lies in the approximate linear dependence between certain sets of independent variables. In this situation relatively small perturbations in the data can cause very large changes in the values of one or more of the estimated coefficients. It is clear then that OLS regression is not robust. One way to ameliorate the problem is to study the individual cases (observations) so that influential observations can be identified and their effects observed. Cook and Weisberg have studied this approach extensively.

Given the aforementioned problems it is not suprising that a good deal of attention has been given to alternative methods of performing regression analyses. Some of those involve definitions of "best" fit which do not involve the minimization of the sum of squared residuals. One such scheme minimizes, instead, the sum of the absolute deviations. A SAS procedure, PROC LAV, has been written to carry out the needed calculations. Another class of alternatives to ordinary least squares involves the weighting of observations according to the size of the residual associated with the observation. Of course once the set of observations has been weighted a new regression model emerges and a new set of residuals is defined. The procedure can quite naturally be thought of as iterative in that the determination of a model leads to a set of residuals and the residuals to a new model, etc, until some stopping rule is satisfied. These iteratively reweighted regression techniques can be used with any of several weight functions. For a survey of such techniques see [1].

Another alternative is based on the jackknife. This approach was introduced by Quenouille in 1949 as technique for reducing the bias in serial correlation estimates. The popularization of the technique is due mainly to John Tukey who advocates its use in a wide variety of situations. See [2] for a general description of the technique and some examples of its use in various situations.

## TUKEY'S CONJECTURE: SUBSEQUENT EVIDENCE

Tukey, [3], surmised that certain functions arising from the jackknife process could be treated as approximately independent and identically distributed random variates. As such, they could be manipulated much as one would normal samples to create a statistic with an approximate t-distribution. While it is true that jackknifing may be used to identify outliers and to provide robust estimates, it is also true that a good deal of the inferential usefulness depends upon this Student's-t approximation.

Miller, [4] and [5], showed that Tukey's conjecture was false in certain situations but also established the asymptotic normality of jackknife estimates of functions of regression parameters. Miller also established that his asymptotic results provided reasonable results for small samples in an inverse regression, or calibration problem.

In order to fix some of the ideas discussed above, consider $\hat{\theta}$, an estimate for a parameter $\theta$ based on a sample of size n. Now delete one of the observations, say the j-th, and denote the new estimate $\hat{\theta}_{-j}$. Delete, in turn, each of the observations so that there are n estimates each based on n-1 observations. Define the psuedo-values

$$\tilde{\theta}_j = n\hat{\theta} - (n-1)\hat{\theta}_{-j}$$

for each j=1,2,...,n. The final jackknife estimate is just the average of the psuedo-values,

$$\tilde{\theta} = \frac{1}{n}\sum_{j=1}^{n} \tilde{\theta}_j$$

where the summation is for j=1,2,...,n. The ratio

$$t = \frac{\tilde{\theta} - \theta}{\sqrt{\frac{1}{n(n-1)} \sum_{j=1}^{n} (\tilde{\theta}_j - \tilde{\theta})^2}}$$

is the approximate t-variate discussed above.

Clearly, for all but the smallest of data sets the jackknife estimation of multiple regression coefficients is too cumbersome to be attempted by hand. A SAS macro is documented in [6] which utilizes the MATRIX procedure to produce the results of a jackknife regression analysis. Some associated output is produced by that macro which can be used in the testing of hypotheses, forming of confidence intervals and plotting of regression lines and residuals.

## THE SIMULATION

In order to assess the "t-ness" of the jackknife t-statistics in various regression situations a simulation was devised in which simple and multiple regression estimates were found for various deterministic models in conjunction with two kinds of "noise". The two distributions selected for the noise or error term were the Normal and the Double Exponential. The later is a bit artificial perhaps because of the cusp in the density function but is commonly used in simulation studies where long tailed noise is desired. Tukey, [7], has indicated that jackknifing may not be appropriate in situations where the distribution of the quantity being jackknifed has heavy tails. We point out that the use of the Double Exponential here for the error term does not produce excessively straggling tails in the dist-ribution of the slope coefficients. The X values for the simple linear regression case were 1(1)10 and 1,3,5,6,6,7,8.5,9,10 following [5]. For the X values in the multiple regression simulation we selected 12 values, at random, from among the 48 observations on page 33 of [9].

The SAS macros written by the current authors and referenced earlier in this article are not appropriate for large scale simulations due to problems of efficiency. A more efficient program was developed so that problems of a small size could be replicated the required number of times. That program utilizes a matrix inverse updating step (See [10]) and is available from the authors upon request.

Our investigation of the robustness of the Tukey conjecture is restricted to the situations in which the errors are Normal or symmetric and long-tailed with zero mean. We specifically do not consider the problem of specification error.

In order to verify that certain parts of the program were working properly and to get some idea of the reproducibility of the results of this simulation ordinary least squares slope estimates were generated using Standard Normal noise superimposed on a model with both slope and intercept equal to unity. That choice of "population" parameters is the same as the one used by Miller in [5]. The simulation was repeated using Double Exponential error terms to assess the effect of this departure from the usual assuptions on OLS estimates for later comparison with similar jackknife results. The Double Exponential variates were generated as the difference of two independent Exponential random variables (See [8]) and the Exponential random variables were, in turn, generated as the negative of the natural logarithms of variates with a Uniform distribution on the interval (0,1). Of course a SAS function was used to generate the Uniform random variables.

Tukey's conjecture involves not only the claim that the t approximation is appropriate but that the approximating distribution is the Student's t-distribution with n-1 degrees of freedom (where n is the number of observations). In this simulation we compared the simulated values of the test statistics with the quantiles of several different t-distributions. The comparison is based on probability plots of the tail regions and on a goodness-of-fit statistic in which the sum (in the tail region) of the absolute deviations between the empirical quantiles and the theoretical values, each divided by the observed value was used. In short, the goodness-of-fit statistic is proportional to the average relative error. We make no claims as to the distributional qualities of the statistic except to assert that "good fits" produce small values for the statistic while "bad fits" produce large values.

## RESULTS

First, we will present the results of the "base case" analysis. That is OLS results for Normal noise imposed on the model 1+X where X is 1(1)10. Note that the distribution of the slope coefficients is known to be Student's-t with n-2, in this case 8, degrees of freedom. Since this distribution is symmetric about zero we have "folded" the empirical distribution by taking the absolute value of each generated observation. In this manner the information in the two tails is, in effect, comined. Another way to say the same thing is to state that the effective sample size is doubled. The probability plot (actually a "Q-Q" plot) appears as Figure 1. In that, the empirical results, based on 10,000 replications, are plotted against the quantiles of six t-distribu-

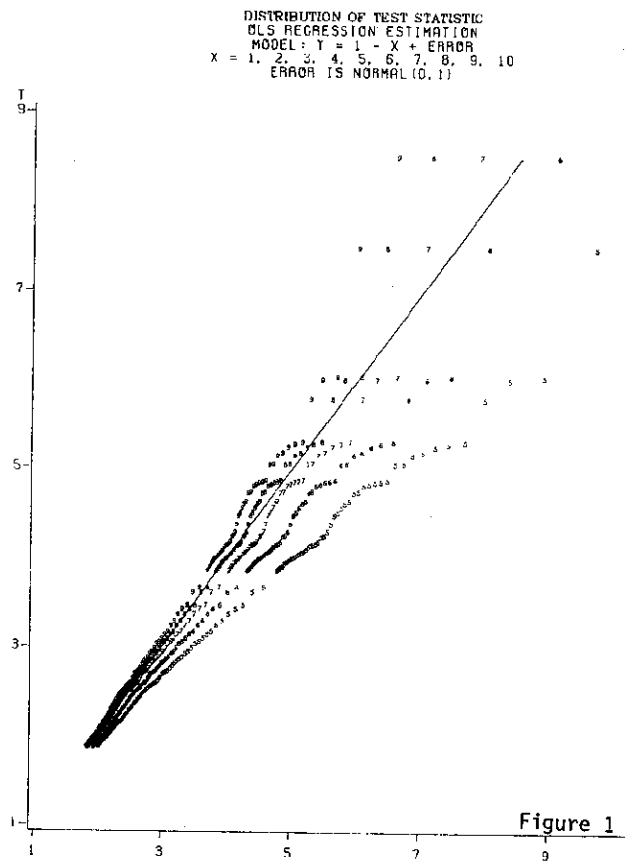


Figure 1

tions (d.f.=5,6,7,8,9,10). In Figure 1 it can be seen that the data fits the eight degree of freedom line rather well except for the largest two values where the lines associated with six and seven degrees of freedom seem to fit better. In terms of the goodness-of-fit statistic mentioned earlier the fit values are

| d.f. | fit |
|---|---|
| 5 | 127.57 |
| 6 | 64.87 |
| 7 | 24.12 |
| 8 | 6.77 |
| 9 | 26.40 |
| 10 | 42.90 |

The eight degree of freedom distribution clearly fits better than the others by this criterion. The critical quantiles of the empirical distribution function and the t-distribution with eight degrees of freedom are matched as follows.

| alpha | empirical | t(8) |
|---|---|---|
| .005 | 3.363 | 3.355 |
| .01 | 2.915 | 2.897 |
| .025 | 2.339 | 2.306 |
| .05 | 1.860 | 1.860 |

* The values used correspond to the fraction i/(10,000+1) closest to the alpha shown.

A similar analysis carried out in the presence of simulated Double Exponential noise produced the results below.

| d.f. | fit |
|---|---|
| 5 | 174.10 |
| 6 | 108.37 |
| 7 | 65.43 |
| 8 | 35.21 |
| 9 | 13.01 |
| 10 | 5.71 |

Subsequent simulations confirm that the value of "fit" ascends for d.f.=11,12,.... The t-distribution corresponding to ten degrees of freedom is seen to yield the best fit. Below the critical quantiles of the "best fit" distribution are compared to the empirical values.

| alpha | empirical | t(10) |
|---|---|---|
| .005 | 3.169 | 3.179 |
| .01 | 2.764 | 2.766 |
| .025 | 2.228 | 2.236 |
| .05 | 1.813 | 1.827 |

In this particular case the usual n-2 degree of freedom test is conservative as may readily be seen in Figure 2.

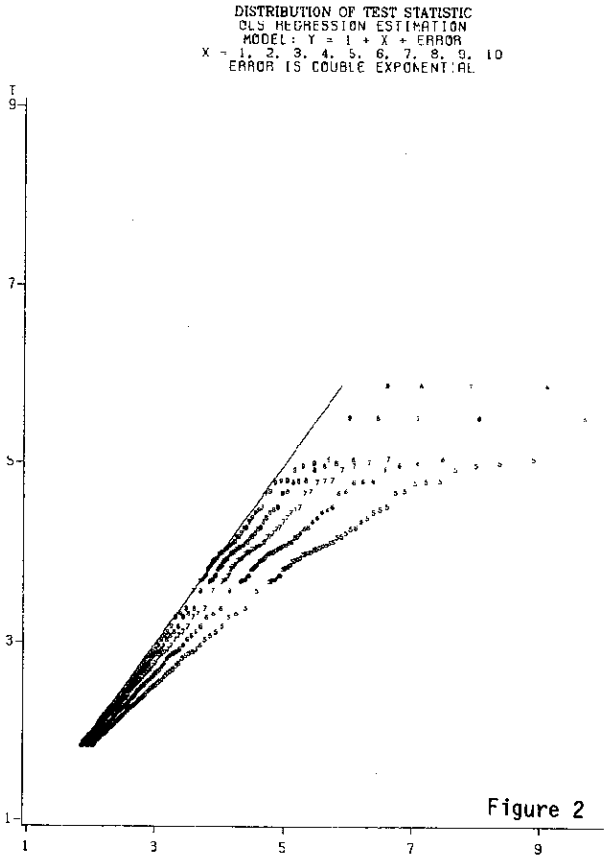


Figure 2

In the case where the x values are bunched up toward the high end, that is for x=1,3,5,6,6,7.8,8.5,9,10 it is known that Normal error terms produce slope

estimates which follow the t-distribution with n-1 degrees of freedom so that situation is not considered. It is not clear, however, that the design points do not affect the distribution of t when the error terms are of the long tailed Double Exponential variety. Our simulations indicate that the opposite may be true as may be seen below.

| d.f. | fit |
|---|---|
| 5 | 162.35 |
| 6 | 97.43 |
| 7 | 55.06 |
| 8 | 25.41 |
| 9 | 13.49 |
| 10 | 16.70 |

Here the nine and ten degree of freedom distributions seem to fit about equally well. A comparison of the empirical results with the critical quantiles of both distributions is given below.

| alpha | empirical | t(9) | t(10) |
|---|---|---|---|
| .005 | 3.170 | 3.250 | 3.169 |
| .01 | 2.750 | 2.822 | 2.764 |
| .025 | 2.259 | 2.262 | 2.228 |
| .05 | 1.856 | 1.833 | 1.813 |

See Figure 3 for graphical representation of the results. Note also that in Figures 1, 2 and 3 the last few observations are somewhat detached from the rest of the observations and also that for the theoretical distributions, especially t(5), the quantiles associated with those observations tend to be quite large. In

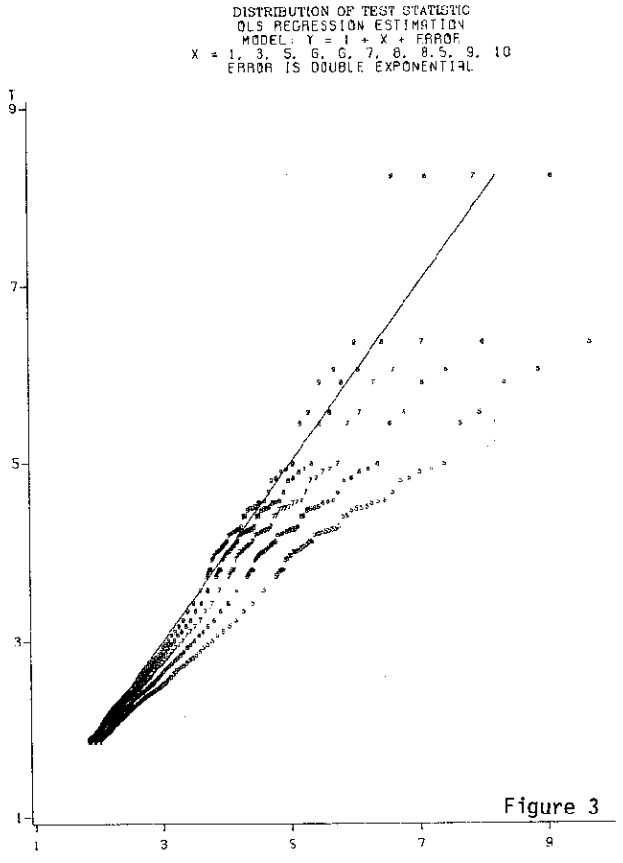

Figure 3

subsequent figures (Figures 4-15) the range of values for both axes are specified in such a way as to eliminate the very extreme observations and quantiles. That is our first, and last, attempt at robust graphics.

Jackknife estimates were calculated for 10,000 replications for the design points and error distributions above in order to assess Tukey's conjecture. In the table below notice that even for Normal noise and X ranging over 1,2,...,10 the "best" t-distribution does not fit nearly as well as it did in the OLS case. We continued to utilize "folding" as discussed in an earlier section. We justified that on the basis of a preliminary test for symmetry (about 0) due to Gupta which is described in [11]. The value of the test statistic, J, turned out to be -0.80 which is nowhere near significance when compared to the reference Normal distribution. The sample size used in that test was 500.

| d.f. | fit |
|---|---|
| 5 | 121.27 |
| 6 | 63.67 |
| 7 | 33.99 |
| 8 | 25.34 |
| 9 | 31.14 |
| 10 | 46.41 |

Note also that t(8) seems to fit better than t(9). Below the empirical quantiles are compared to the t-distributions with 7,8 and 9 degrees of freedom.

| alpha | empirical | t(7) | t(8) | t(9) |
|---|---|---|---|---|
| .005 | 3.584 | 3.500 | 3.355 | 3.250 |
| .01 | 2.974 | 2.998 | 2.897 | 2.822 |
| .025 | 2.282 | 2.365 | 2.306 | 2.262 |
| .05 | 1.819 | 1.895 | 1.860 | 1.833 |

It is clear that no single t-distribution does a very good job of fitting the four quantiles given here. The six degree of freedom distribution provides a conservative rule in these four cases but is perhaps too much so in some of the cases. See Figure 4 for the probability plots.

The simulation which involved the long tailed error terms and the evenly spaced X values yields results which are markedly different from those of the previous paragraph. See Figure 5 for the graphical representation.

| d.f. | fit |
|---|---|
| 5 | 204.51 |
| 6 | 137.66 |
| 7 | 93.94 |
| 8 | 63.16 |
| 9 | 41.59 |
| 10 | 29.70 |

It is not clear if a better fit might have derived from t(11) or t(12) although Figure 5 suggests that is probably not the case. The comparison of quantiles (not shown here) indicates that t(10) approximates the quantiles associated with significance levels of .005 and .01 very well

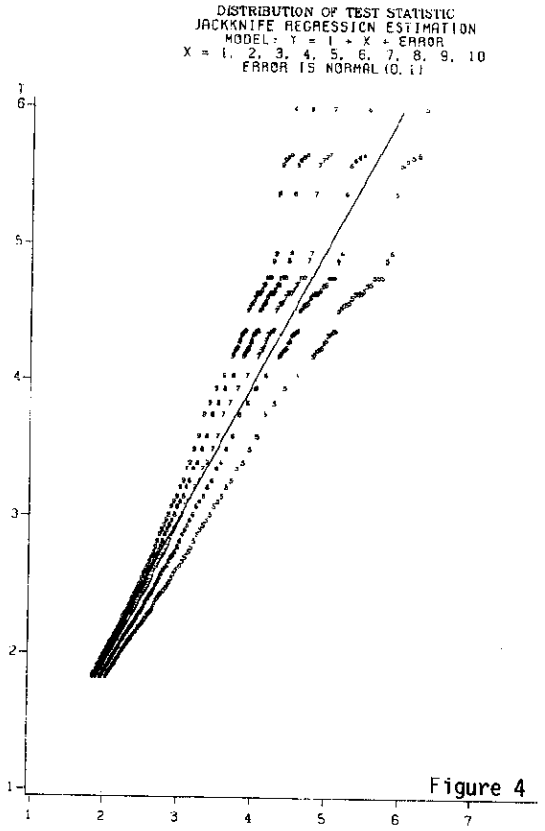


Figure 4

but is a little conservative beyond that point.

For the design points which are bunched up toward the larger values similar results appear. For Double Exponential noise the results are as follows.

| d.f. | fit |
|---|---|
| 5 | 149.87 |
| 6 | 88.50 |
| 7 | 50.77 |
| 8 | 37.02 |
| 9 | 34.92 |
| 10 | 37.88 |

Clearly, there is not much reason to prefer any one of the three best fitting distributions to any of the others. The the quantiles of the t-distribution with 8 degrees of freedom (perhaps even the one with 7) seem to fit the quantiles associated with significance levels of .005 and .01 but the more extreme quantiles are badly overestimated by those distributions. In this case the tails of the empirical distribution seem to have a different shape than those found in the family of t-distributions. The results for a Normally distributed error term are quite a bit different. The fit values are given below.

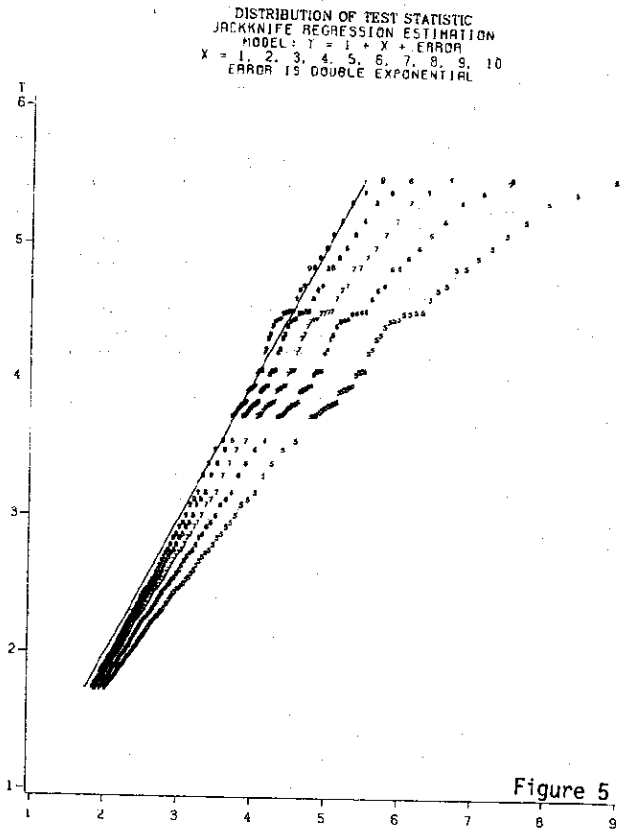


Figure 5

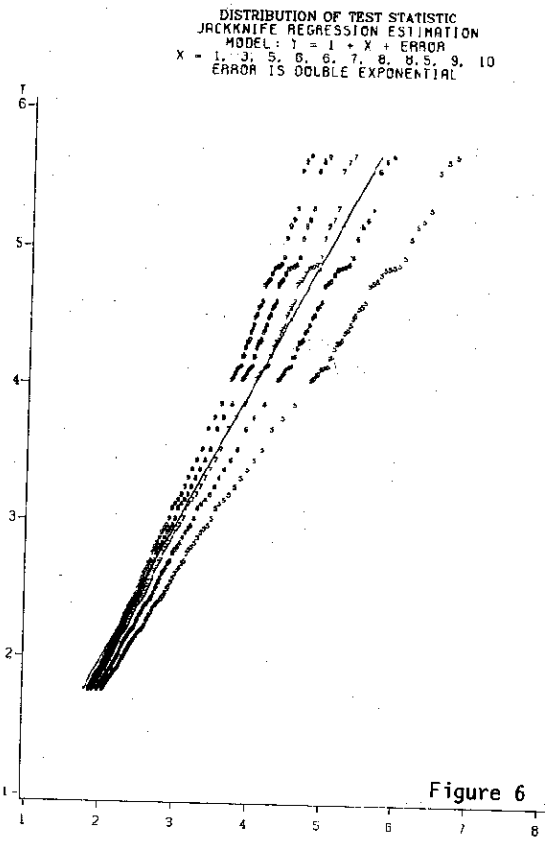


Figure 6

| d.f. | fit |
|---|---|
| 5 | 43.18 |
| 6 | 18.58 |
| 7 | 51.96 |
| 8 | 78.27 |
| 9 | 97.79 |
| 10 | 112.85 |

The best fit is given by t(6). The comparison of the emipical quantiles to those of t(6) follows.

| alpha | empirical | t(8) |
|---|---|---|
| .005 | 3.802 | 3.708 |
| .01 | 3.208 | 3.143 |
| .025 | 2.479 | 2.447 |
| .05 | 1.925 | 1.943 |

Exceptin the most extreme case t(6) approximates the quantiles very well. We note, however that the best fit corresponds to a t-distribution fewer degrees of freedom than that proposed by Tukey.

For our multiple regression simulations we used as a guide a data set and analysis which appears in [9]. The data set contains measurements of the dependent variable, Y, (per capita fuel consumption) and several related independent variables. The data set contains 48 observations, one for each of the contiguous states. Twelve of the observations were chosen at random to provide the values for the independent variables in this simulation. The dependent variable was generated by using Weisberg's parameter estimates for the entire data set in conjunction with the independent variables and then adding some error term to that result. The variance of the error term was set equal to the variance estimate from the full data set. Both Normal and Double Exponential noise was used. Jackknifing was then used to estimate each of the four slope coefficients. Figures 8-11 contain the probability plots for the situations when Gaussian noise was added and Figures 12-15 reflect the results for the long tailed error terms. In the table which follows we present the goodness of fit statistics for each of the four independent variables using Normal noise.

| d.f. | TAX | DLIC | INC | ROAD |
|---|---|---|---|---|
| 6 | 208.00 | 227.09 | 202.14 | 127.29 |
| 7 | 161.64 | 180.18 | 157.50 | 84.50 |
| 8 | 129.00 | 147.15 | 126.44 | 55.32 |
| 9 | 104.88 | 123.08 | 105.57 | 37.90 |
| 10 | 86.80 | 105.76 | 90.69 | 32.96 |
| 11 | 74.39 | 92.87 | 79.99 | 32.15 |
| 12 | 65.69 | 83.72 | 73.75 | 33.50 |

It would seem that the coefficients for the variable ROAD more closely follow a Student's-t distribution than do the other coefficients. Subsequent runs indicate that t(13) fits two of the empirical distributions better than any of the distributions above. In any event it

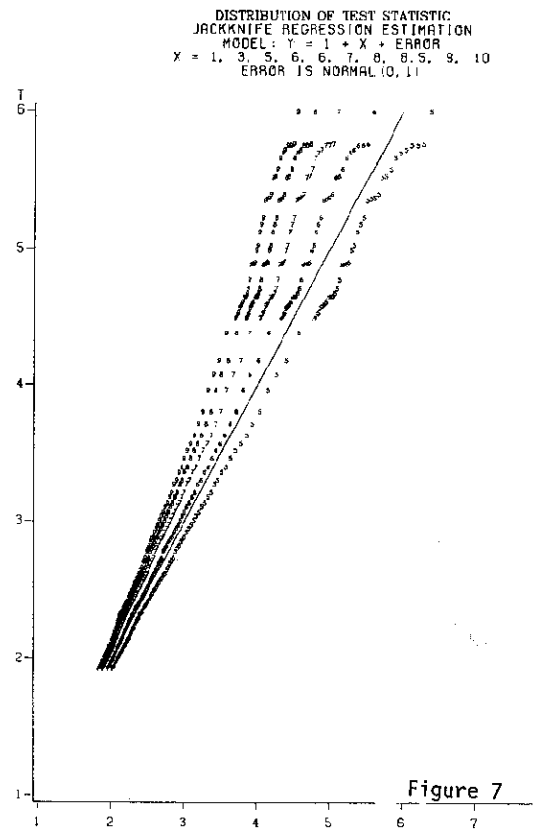


Figure 7

would seem that the t approximation is not particularly accurate in this situation. An examination of the observed quantiles indicates that Tukey's n-1=9 degree of freedom rule is conservative in all 16 of the cases (4 variables times 4 quantiles). In fact, it appears that a test using the t-distribution with 10 degrees of freedom is conservative in 15 of the 16 cases. Even though the evidence indicates that the t approximation is not very good it is true, apparently, that tests of hypothesis andconfidence intervals based on t(9) are valid(if conservative).

The multiple correlation simulation done with Double Exponential error terms yields results similar the previous case. Again the goodness of fit statistics do not seem to indicate very good fits for any of the seven t-distributions which we tried. Again the n-1 degree of freedom rule appears to be conservative in all instances, in fact even t(10) appears to give conservative results in all 16 cases. For this situation even Millers' Normal approximation appears to give results which are either conservative or slightly anticonservative for significance levels of .005 and .01. Apparently long tailed error terms have the effect of producing fewer extreme t values than do Gaussian error terms. It would interesting to see if error distributions with abrupt ends result in long tails for the distribution of the slope estimates.

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS NORMAL (0, 66.3)

DATA FROM TABLE 2.1, APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR TAX COEFFICIENT

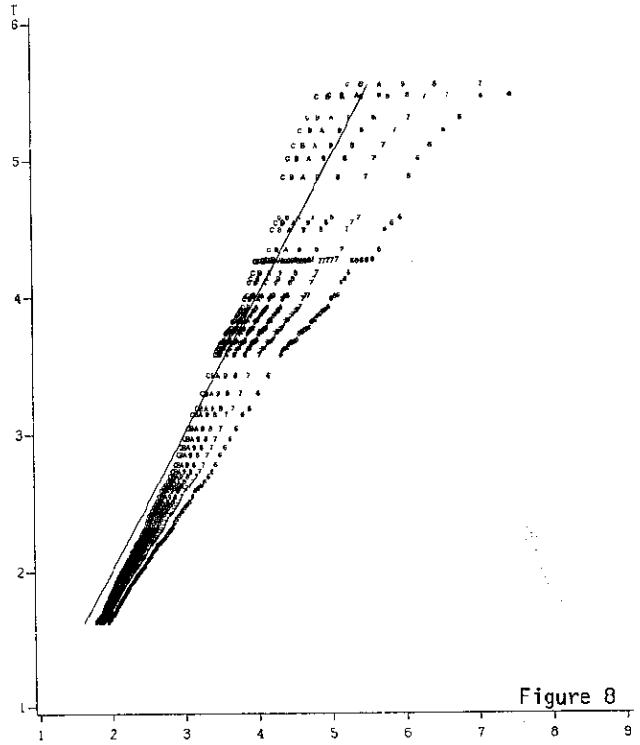

Figure 8

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS NORMAL (0, 66.3)

DATA FROM TABLE 2.1, APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR INC COEFFICIENT

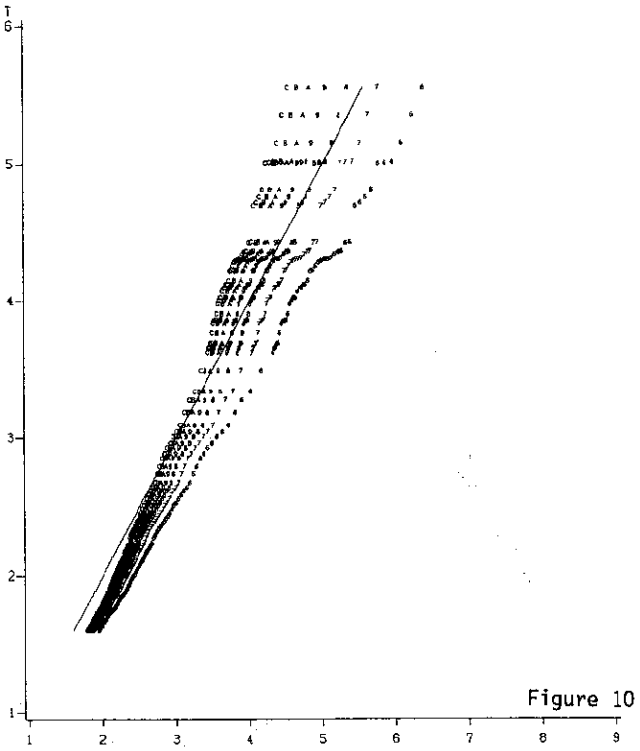

Figure 10

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS NORMAL (0, 66.3)

DATA FROM TABLE 2.1, APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR DLIC COEFFICIENT

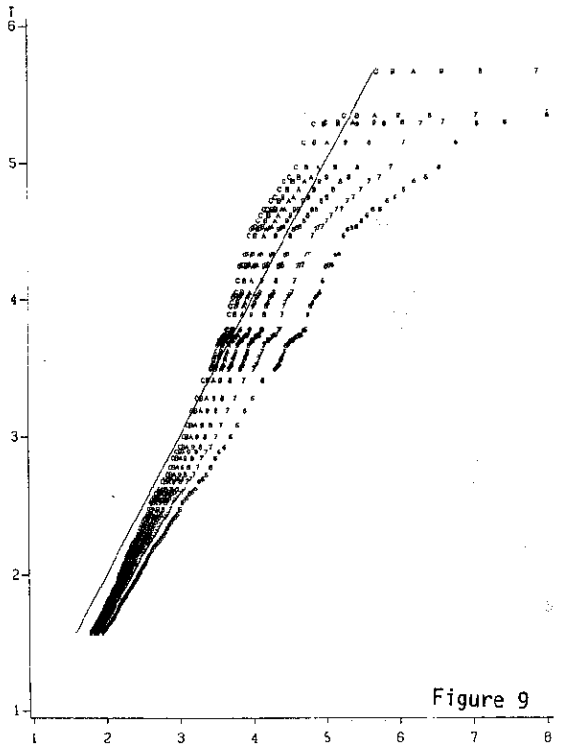

Figure 9

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS NORMAL (0, 66.3)

DATA FROM TABLE 2.1, APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR ROAD COEFFICIENT

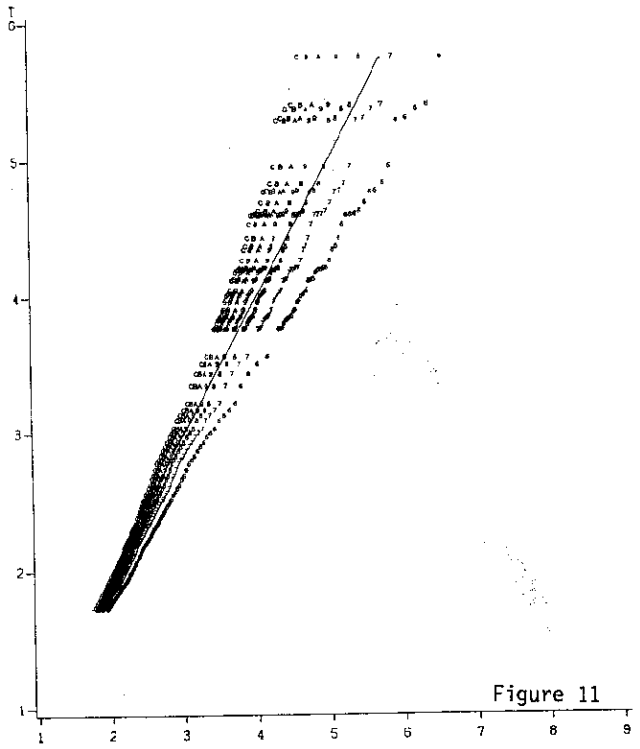

Figure 11

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS DOUBLE EXPONENTIAL

DATA FROM TABLE 2.1. APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR TAX COEFFICIENT

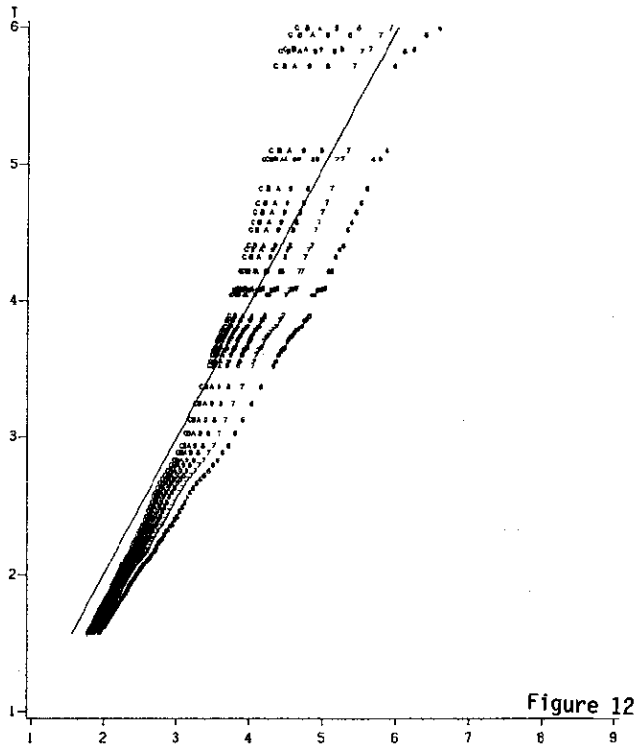

Figure 12

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS DOUBLE EXPONENTIAL

DATA FROM TABLE 2.1. APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR INC COEFFICIENT

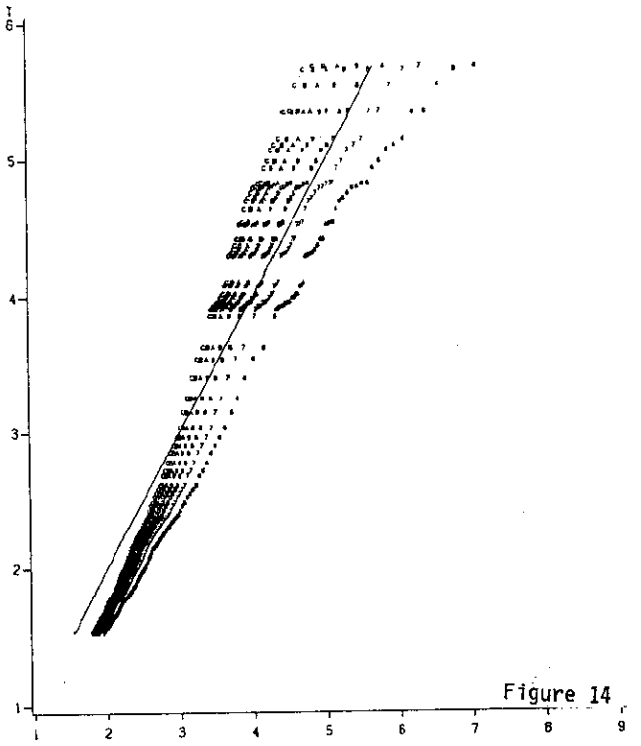

Figure 14

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS DOUBLE EXPONENTIAL

DATA FROM TABLE 2.1. APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR DLIC COEFFICIENT

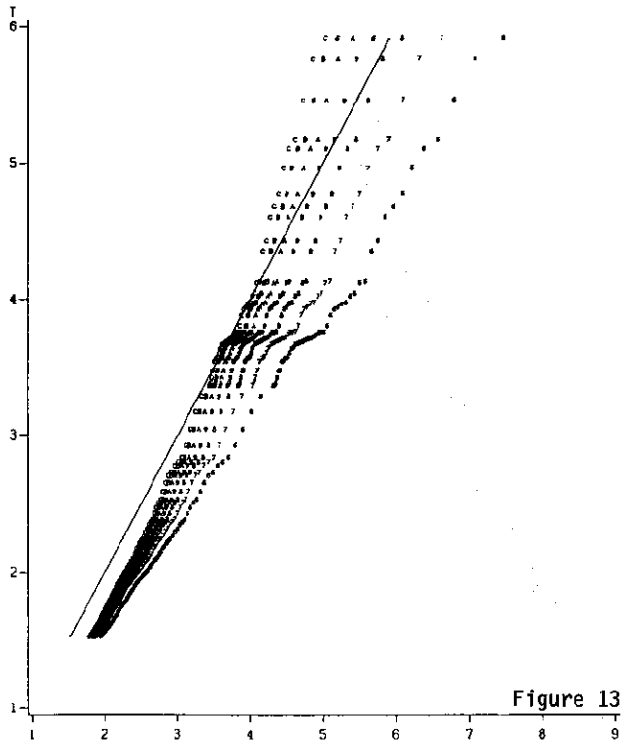

Figure 13

DISTRIBUTION OF TEST STATISTIC
JACKKNIFE MULTIPLE REGRESSION
MODEL: Y = 377.3 - 34.8*TAX + 1336.45*DLIC - .0666*INC - .00243*ROAD
ERROR IS DOUBLE EXPONENTIAL

DATA FROM TABLE 2.1. APPLIED LINEAR REGRESSION BY WEISBURG

TEST FOR ROAD COEFFICIENT

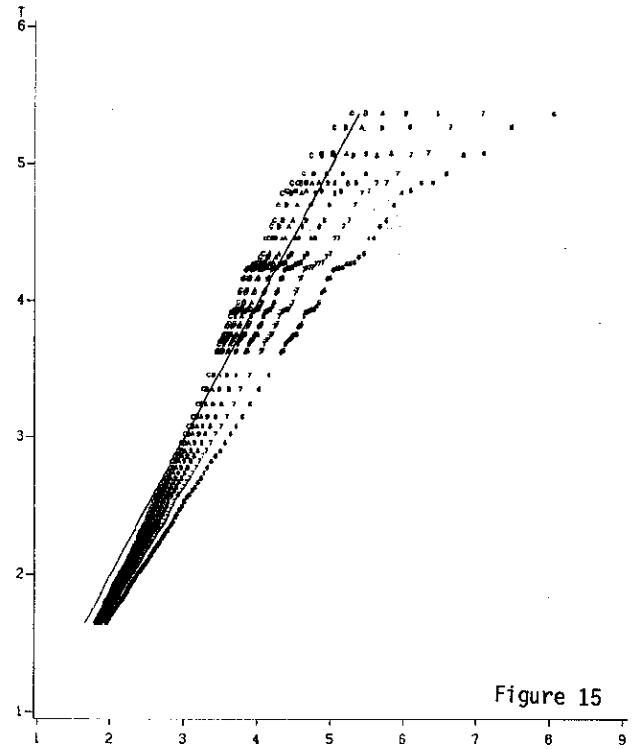

Figure 15

## LITERATURE CITED


[1] Hogg, R.V., Statistical Robustness: One View of its Use in Applications Today, The Amer. Statistician, Vol 33-3 (1979)

[2] Mosteller & Tukey, Data Analysis and Regression, Addison-Wesley (1977)

[3] Tukey, J.W., Bias and Confidence in not-quite large samples, Ann. Math. Stat. Vol 29 (1958)

[4] Miller, R.G., A Trustworthy Jackknife, Ann. Math. Stat., Vol 35 (1964)

[5] _____, An Unbalanced Jackknife, Ann. of Stat., Vol 2 (1974)

[6] Hobbs, G.R. et al, Jackknife Techniques in Regression Analysis, SUGI Vol 6 (1980)

[7] Tukey, J.W., Data Analysis and Behavioral Science (unpublished) Princeton University

[8] Bickel & Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Holden-Day (1976)

[9] Weisberg, S., Applied Linear Regression, Wiley (1980)

[10] Duncan & Horn, Linear dynamic recursive estimation from the viewpoint of regression analysis, JASA, Vol 67 (1972)

[11] Hollander & Wolfe, Nonparametric Statistical Methods, Wiley (1973)