Staff Paper P89-19 June 1989

Staff Papers Series

CHOICE OF REGRESSION METHOD

FOR DETRENDING TIME SERIES DATA

WITH NONNORMAL ERRORS

by
Scott M. Swinton and Robert P. King

Department of Agricultural and Applied Economics

University of Minnesota
Institute of Agriculture, Forestry and Home Economics
St. Paul, Minnesota 55108

Staff Paper P89-19 June 1989

CHOICE OF REGRESSION METHOD FOR DETRENDING TIME SERIES DATA WITH NONNORMAL ERRORS*

by
Scott M. Swinton and Robert P. King**

Department of Agricultural and Applied Economics
University of Minnesota
St. Paul, MN 55108

* Selected paper to be presented at the annual meeting of the American Agricultural Economics Association, Louisiana State University, Baton Rouge, Louisiana, July 30-August 2, 1989.

** Scott M. Swinton is a graduate research assistant and Robert P. King is a professor in the Department of Agricultural and Applied Economics, University of Minnesota, St. Paul, Minnesota.

Introduction

In studies of probability distributions for variables which are observed sequentially through time (e.g., production by an individual firm), it is often necessary to detrend data in order to eliminate bias due to changes in such factors as technology and tastes. This is generally accomplished by regressing the dependent variable on some measure of time using least squares methods. These give maximum likelihood estimates under the assumption that the error term is normally distributed. While this may be an attractive simplifying assumption, empirical tests have found that it often fails to hold for agronomic and farm-level crop yield data (Day). Moreover, even for data that are distributed normally, measurement and input errors commonly "contaminate" the data set.

Robust regression (RR) techniques offer means of coping with data for which the error term is not normally distributed. They have been discussed extensively in the statistics literature during the past 20 years and are becoming increasingly available to applied economists in econometric software packages (e.g., SHAZAM (White et al.), PROGRESS (Rousseeuw and Leroy)). However, they do not necessarily offer coefficient estimates that are significantly different from ordinary least squares (OLS), even when errors are not distributed normally. This paper will (1) illustrate cases in which several robust regression methods on an equation having nonnormal errors failed to give coefficient estimates significantly different from OLS and

(2) demonstrate how regression diagnostics can identify conditions under which RR is a useful alternative to OLS for detrending time series data.

Nonnormal errors and robust regression

In the standard linear model,

$$Y = X'\beta + u,$$

OLS will yield maximum likelihood estimates of model parameters if the error term, **u**, is an independently, identically distributed normal random variable with mean 0 and variance σ^2. When the error term is not distributed normally, the OLS estimator, **b**, is still the best linear unbiased estimator, and the OLS variance estimator, s^2, is still unbiased and consistent. However, neither one is efficient or asymptotically efficient, since the maximum likelihood estimator is nonlinear (Judge et al. 1988, p. 888). As a general-purpose alternative, robust statistics have been proposed that offer a "lack of susceptibility to the effects of nonnormality" (Mosteller and Tukey, p. 16), while still offering relatively efficient estimates when errors are, in fact, normally distributed.

In recent years, two schools of thought have emerged on how to handle nonnormal errors when they arise in regression analyses: the regression diagnostics school and the robust regression school. While both seek to find the best model for the data when distorting outliers may be present, their approaches are quite different. The key differences are aptly summarized by Rousseeuw and Leroy (p. 75):

Regression diagnostics first attempt to identify points that have to be deleted from the data set, before applying a regression method. Robust regression tackles these problems in the inverse order, by designing estimators that dampen the impact of points that would be highly influential otherwise.

Robust regression methods give less weight to observations which deviate far from the expected value of the dependent variable than does OLS. They are classified by their approach to controlling infuential outliers. M-estimators employ maximum-likelihood techniques for finding regression coefficients that will minimize some function of the regression residuals, typically a function that down-weights residuals large in absolute value. Linear combinations of order statistics, or L-estimators, calculate regression coefficients for quantiles of the residuals resulting from a regression model and then combine them with specified weights. Rank, or R-estimators, constitute a third category of robust estimator, this based on minimization of a sum of ranks weighted by corresponding scores. For a survey of RR methods, see Judge et al. 1985, Huber 1977 or Huber 1981.

Robust regression: Not a panacea for nonnormal errors

Although robust regression techniques are recommended for obtaining estimates that are more efficient than OLS; the coefficient estimates will not necessarily be different. This is illustrated by results from six robust regression methods on three samples of time series corn yield data. The data came from three farms in Jackson County, Minnesota, having 15 to 43 years of observations. All series end between 1985 and 1987. Analysis of OLS residuals showed evidence

of negative skewness. This finding confirmed results of King and Benson using similar data in an earlier study.

Due to the nonnormality of the OLS residuals, detrending regressions were performed for each farm using RR methods on the linear model

CORNYLD=f(constant,YEAR)

where CORNYLD denotes the corn grain yield in bushels per acre and YEAR is the corresponding year. Six different RR methods discussed in Judge et al. (1988) were considered: the multivariate t M-estimator and five L-estimators, the least absolute error (LAE), trimmed mean (TRIM), five quantity weighted regression quantile (FIVEQUAN), the Gastwirth weighted regression quantile (GASTWIRTH), and Tukey tri-mean weighted regression quantile (TUKEY) procedures. These were implemented on the SHAZAM version 6.0 econometrics package (White et al.).

As summarized in Table 1, all estimates of the coefficient on YEAR lay within one standard error of the estimates from OLS estimates. Coefficient estimates with RR were insignificantly different from OLS despite (1) a disproportionate number of large negative residuals, (2) negative summed residuals for all RR methods used, and (3) a very different coefficient estimate for Farm C from those of farms A and B.

Regression diagnostics

Regression diagnostics offer some explanations for the special case of a regressor that is a series with unit increments. Two common regression diagnostic measures are particularly helpful: (1) the measure of potential leverage to influence the regression individual

Table 1: Coefficient estimates compared for YEAR variable in the linear model: OLS versus six robust regression methods.

Regression method	Farm A	Farm B	Farm C
Parameters			
Degrees of freedom (d.f.)	40	41	13
OLS standard error of coef.	0.214	0.242	1.154
Coefficient estimates			
OLS	2.034	2.062	6.442
LAE	2.166	2.203	6.943
TRIM=.05	2.039	2.209	--[1]
TRIM=.10	2.049	2.184	--[1]
TRIM=.20	2.129	2.086	7.220
FIVEQUAN	2.073	2.033	--[1]
GASTWIRTH	2.068	2.083	6.929
TUKEY	2.108	2.091	6.226
MULTIT=1	2.133	2.149	6.766
MULTIT=3	2.097	2.127	6.671
MULTIT=d.f.	2.042	2.075	6.527

[1] These L-estimates could not be computed due to small sample size relative to the size of the desired trim quantile.

observations on the independent variables, h_{ii}, and (2) the measure of what constitutes an influential outlier, the "studentized" residual.

The potential influence, or leverage, of an observation, h_{ii} (Belsley et al., Judge et al. 1988, Weisberg), is defined as follows,

$$h_{ii} = x_i'(X'X)^{-1}x_i,$$

where x_i is an observation of the independent variable(s) and X is the matrix of all observations on the independent variable(s). Key characteristics of h_{ii} are (1) it always lies between zero and one, (2) the h_{ii}'s sum to k, the number of regressors, and (3) h_{ii} is a function of the independent variables only. As observations on the independent variables, x_i, get farther from the sample mean, they become potentially more influential, and h_{ii} grows larger. Finally, there is an inverse relationship between the h_{ii} and the sample variance (Weisberg, p. 110), since

$$\text{var}(e) = \sigma^2(1-h_{ii}).$$

Since the average value of h_{ii} is k/n, a conservative rule of thumb for observations with high leverage is $h_{ii} \geq 3k/n$ (Judge et al. 1988, p. 893).

A widely employed, reliable measure of whether an extreme residual is indeed a statistical outlier is the "studentized" or "externally studentized" residual (Belsley et al., Weisberg), e_i^*,

$$e_i^* = \frac{e_i}{s(i)^*(1-h_{ii})^{.5}},$$

where e_i is the residual corresponding to the i^{th} observation and s(i) denotes the sample standard error of the estimate calculated omitting the i^{th} observation. As is clear from its composition, e_i* will be large if one or more of the following conditions obtains: (1) e_i is large, (2) h_{ii} is large, or (3) s(i) is small.

The studentized residual follows the central t-distribution with n-k degrees of freedom. However, since it describes a residual which represents one of many "draws" from the distribution, the appropriate test statistic is the Bonferroni t-value, t_i, which tests the hypothesis that the residual would be likely to occur with a/n probability, where a is the probability of mistakenly rejecting the hypothesis that e_i is an outlier, and n is the number of observations in the sample (Weisberg, p. 116).

Belsley et al. have developed a statistic called DFBETAS to measure the influence that an observation is likely to have on the regression coefficient. It measures the difference between estimates for the j^{th} coefficient with and without the i^{th} observation as standardized by the corresponding coefficient standard error (in the denominator) (p. 13):

$$DFBETAS_{ij} = \frac{b_j - b_j(i)}{s(i)[(X'X)_{jj}^{-1}]^{.5}}.$$

Belsley et al. demonstrate that DFBETAS decreases with sample size at a rate proportionate to $n^{.5}$. Hence, they recommend a "size-adjusted cutoff" of $|DFBETAS| > 2/(n^{.5})$ (p. 28).

Diagnosis of residuals from regressions to detrend time series

A regression to detrend time series is a simple regression in which the independent variable is a unit of time. In uninterrupted time series, the time measure will increase by a single unit from one observation to the next. Even in interrupted time series, with economic data we tend to encounter fairly small breaks in the series. This property of time series data has distinct consequences for regression diagnostics.

First, the measure of potential leverage, h_{ii}, cannot become very large because there are no x_i values far from the mass of x_i's. The h_{ii} values are greatest at the beginning and end of the series, following a symmetric U-shaped pattern of decrease from the starting point to the mean/median value and increase up to the end point. Since the h_{ii}'s sum to the number of regressors, in a simple regression they sum to 2. Hence, as the number of sequential observations increases, the potential leverage of any one decreases (Table 2). Note that using the conservative Judge et al. (1988) cutoff value of 3k/n, none of the h_{ii} values in the table indicate that the observation appears to have unusual potential influence.

Since h_{ii} values are constrained from becoming especially large in regressions detrending a single time series, large studentized residuals can occur only if (1) **s(i)** is small, (2) e_i is large, especially if, in addition, (3) **n** is small (making h_{ii} larger). Note that the first two conditions are not likely to obtain if **n** is very small, since the fitted regression minimizes the e_i^2. Moreover, given

Table 2: Potential leverage coefficients, h_{ii}, by observation for single time series samples ranging from 3 to 40 observations.

Observation number	Number of observations in sample						
	3	5	10	15	20	30	40
1	.833	.600	.346	.242	.186	.127	.096
2	.333	.300	.249	.195	.159	.114	.089
3	.833	.200	.176	.156	.135	.103	.083
4		.300	.127	.124	.114	.092	.076
5		.600	.103	.099	.096	.082	.070
6			.103	.081	.081	.074	.064
7			.127	.070	.068	.066	.059
8			.176	.067	.059	.058	.054
9			.249	.070	.053	.052	.050
10			.346	.081	.050	.047	.046
11				.099	.050	.042	.042
12				.124	.053	.039	.039
13				.156	.059	.036	.036
14				.195	.068	.034	.033
15				.242	.081	.033	.031
16					.096	.033	.029
17					.114	.034	.027
18					.135	.036	.026
19					.159	.039	.025
20					.186	.042*	.025*

* Values continue symmetrically to 30 and 40 observations, respectively.

the routine variability of much agricultural data, **s(i)** and e_i often do not reach orders of magnitude great enough to become significant.

The most reliable regression diagnostic for revealing the influence of an observation on coefficient estimates is DFBETAS. This became especially evident in a test of regression diagnostics and robust methods on a "contaminated" data set.

The test entailed regressions on corn yield data from a southwest Minnesota farm for various series of 5 to 40 years up to 1984, except that the 1984 value was replaced by a yield three standard errors below the expected yield of 122.5 bushels per acre (72.7 bu/ac). The test was conceived to model the effect that might be expected from low yields caused by the 1988 drought. Table 3 compares regression diagnostics from the 1984 observation and presents estimates of the coefficient on YEAR. Of the three regression diagnostic measures, only DFBETAS signalled a likely problem with the 1984 observation. The h_{ii} values all remained below the cutoff value, as shown in Table 2. The studentized residuals were small for small samples, because the large outlier biased the regression, reducing e_i and increasing **s(i)**. None of the studentized residuals exceeded the Bonferroni critical t-value, which is dependent on sample size.

Although robust regression has been recommended as a pre-diagnostic technique (Weisberg p. 253), Table 3 demonstrates that robust methods are not foolproof. Only the 20 percent trimmed mean generated coefficients consistently within two coefficient standard errors of the OLS estimate on the uncontaminated sample (1.88 ± 0.23). However, trimmed mean estimates cannot be computed for sample sizes

Table 3: Comparison of regression diagnostics and robust regression methods on corn grain yield detrending regressions for Farm A with a "contaminated"[1] 1984 observation.

Number of observ.	Year series	Studentized residual for 1984	DFBETAS	Coefficient on YEAR			
				OLS	LAE	Mt=1[2]	TRIM=.2
5	1980-84	-1.02	-1.02+	-14.32 (5.24)	--	-17.16	--
10	1975-84	-2.30	-1.41+	1.13 (2.96)	4.61	2.46	1.87*
15	1970-84	-2.10	-1.01+	1.05 (1.44)	2.59	1.41	1.55*
20	1965-84	-2.52	-1.03+	2.22 (0.89)	2.67	2.56	1.86*
25	1960-84	-2.53	-0.92+	1.93* (0.58)	2.67	2.26*	2.00*
30	1955-84	-2.68	-0.88+	1.91* (0.41)	2.51	2.23*	2.07*
35	1950-84	-2.61	-0.79+	1.84* (0.32)	2.51	2.13*	2.22*
40	1945-84	-2.63	-0.74+	1.73* (0.25)	1.96*	1.92*	1.82*

N.B.: Standard error in parentheses.

1 The "contaminated" 1984 observation was 3 standard errors of estimate below the expected value for that year.

2 Denotes Multivariate t distribution with 1 degree of freedom.

+ DFBETAS value exceeds size adjusted cutoff of $2/(n^{.5})$.

* Coefficient estimate lies within two coefficient standard errors of uncontaminated estimate for 1945-84, 1.88 ± 0.23.

too small to allow at least two observations in the quantiles to be cut from each tail. The least absolute errors (LAE) estimator performed very poorly, and the M-estimator following the multivariate t distribution with one degree of freedom did no better than the OLS estimator. While a significant difference between coefficient estimates with OLS and with RR may be cause for examining residuals, lack of a difference between the two (at least as measured by the OLS coefficient standard error) is not sufficient reason for complaisance.

Conclusion

Cases used to illustrate the value of robust regression methods in simple regression draw on data which has outliers among the observations on the independent variable (cf. Rousseeuw and Leroy, Hampel et al.). However, when the independent variable is a measure of time, such outliers are rare. In the instance of an extreme value of the dependent variable at the end of a data series, it has been demonstrated that robust techniques may fail to outperform OLS. In such a case, the only reliable indicator that an individual observation has a significant impact on coefficient estimates is DFBETAS (or a similarly constructed diagnostic statistic). Unfortunately, this diagnostic is cumbersome to consult when the sample size is large and/or the number of regressors large. As sample size grows, the potential leverage of even an end-of-series outlier decreases, so OLS improves in reliability.

REFERENCES CITED

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley.

Day, Richard H. 1965. "Probability Distributions of Field Crop Yields." Journal of Farm Economics 47(3):713-741.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley.

Huber, Peter J. 1977. Robust Statistical Procedures. CBMS-NSF Regional Conference Series in Applied Mathematics No. 27. Society for Industrial and Applied Mathematics (Philadelphia).

-------. 1981. Robust Statistics. Wiley.

Judge, George G., W.E. Griffiths, R. Carter Hill, Helmut Lutkepohl, and Tsoung-Chao Lee. 1985. The Theory and Practice of Econometrics. Second edition. Wiley.

Judge, George G., R. Carter Hill, W.E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee. 1988. Introduction to the Theory and Practice of Econometrics. Second edition. Wiley.

King, Robert P. and Fred J. Benson. 1985. "Summary Report of Minnesota Crop Insurance Study." Unpublished research report. Department of Agricultural and Applied Economics, University of Minnesota, St. Paul.

Mosteller, Frederick and John W. Tukey. 1977. Data Analysis and Regression. Addison-Wesley.

Rousseeuw, Peter J. and Annick M. Leroy. 1987. Robust Regression and Outlier Detection. Wiley.

Weisberg, Sanford. 1985. Applied Linear Regression. Second edition. Wiley.

White, Kenneth J., Shirley A. Haun and Nancy G. Horsman. 1987. SHAZAM: The Econometrics Computer Program Version 6. Department of Economics, University of British Columbia.