How Robust Is Linear Regression with Dummy Variables ?

Eric Blankmeyer
Department of Finance and Economics
McCoy College of Business Administration
Texas State University
San Marcos, TX 78666

Email eb01@txstate.edu
Tel. 512-245-3253

November 2006

Abstract

Researchers in education and the social sciences make extensive use of linear regression models in which the dependent variable is continuous-valued while the explanatory variables are a combination of continuous-valued regressors and dummy variables. The dummies partition the sample into groups, some of which may contain only a few observations. Such groups may easily include enough outliers to break down the parameter estimates. Models with many fixed or random effects appear to be especially vulnerable to outlying data. This paper discusses the problem at an intuitive level and cites sources for the key theorems establishing bounds on the breakdown point in models with dummy variables.

Key words. Breakdown point, outliers, fixed effects

1. INTRODUCTION

Researchers in education and the social sciences make extensive use of linear regression models in which the dependent variable is continuous-valued while the explanatory variables are a combination of continuous-valued regressors and dummy (binary) variables. The role of the dummies is to partition the data set into two or more groups based on qualitative criteria. However, the inclusion of dummies tends to degrade the robustness of linear regression estimators when the sample contains anomalous observations. Statisticians have been aware of this problem for at least a decade (e. g. Mili and Coakley 1996), but researchers in other fields seem not to have focused on it. Their panel-data models, often heavily parameterized with fixed effects, are potentially quite vulnerable to atypical data. This paper discusses the problem at an intuitive level and cites key theorems on the breakdown point of linear regression with dummy variables. Robust estimation of these models also raises computational issues, a topic that has been examined by Hubert and Rousseeuw (1997) and by Maronna and Yohai (2000).

2. THE BREAKDOWN POINT AND EQUIVARIANCE

Observations are not anomalous per se, but only with respect to a particular regression model. This means, for example, that an outlier may not be stand out among the observations on the dependent variable or among the observations on any particular regressor; the datum is simply at odds with the economic and statistical assumptions of the researcher’s linear model. If the

anomalies cannot be resolved, it will usually be prudent to remove those observations from data set. Of course, outliers should not be downweighted or discarded automatically; they sometimes indicate that the model itself needs to be revised. A robust regression method supports this kind of diagnosis because it does not hide the outliers but actually reveals them as large residuals.

An estimator's robustness can be characterized in several ways. One concept that has received a lot of attention in recent years is an estimator's breakdown point (Donoho and Huber 1983; Rousseeuw and Leroy 1987, chap. 1); *it is the smallest fraction of contamination that can produce an infinite bias in the estimator*. For instance, in a univariate sample of size n, the average can be increased without limit if any single observation is made arbitrarily large; accordingly, the breakdown point of the average is 1/n, or zero asymptotically. On the other hand, if all the observations that exceed the sample median are increased arbitrarily, the median is unchanged; so its breakdown point is essentially 50%. In other words, the breakdown point accurately characterizes the average's well-known lack of robustness and the sturdiness of the median.

Contamination in excess of an estimator's breakdown point is a sufficient condition for an indefinitely large bias, but it is not a necessary condition. As a practical matter, the extent of the bias obviously depends not only on the number of outliers but also on their magnitudes. Another important consideration is the fit between the uncontaminated data and the model. In a linear regression where the valid observations have a very high R-squared, the bias induced by several stray data points may be small even if the estimator itself is not highly robust.

Conversely, if the valid observations have a low R-squared, application of a high-breakdown estimator may not avoid a large (but finite) bias due to a few outliers. Accordingly, the breakdown point is similar to estimation criteria like efficiency and consistency; its usefulness ultimately depends on its performance in actual applications (Zaman et al. 2001). While other concepts, such as the influence function (Hampel et al. 1986), can also be used to evaluate robustness, this paper focuses on the breakdown point because it greatly clarifies the impact of dummy variables on a regression when the sample is contaminated.

Another important property of an estimator is equivariance. A linear regression estimator is equivariant if it transforms properly when a variable (either dependent or regressor) is recentered or rescaled (Rousseeuw and Leroy 1987, p. 116). For example, if each observation on a particular continuous-valued regressor is multiplied by a positive constant c, the estimated regression coefficient should change by the factor 1/c. Some widely-used equivariant estimators are least squares, the L_1 norm (least absolute errors), and least trimmed squares (Rousseeuw and Leroy 1987, pp. 132-135). On the other hand, orthogonal regression is a method that lacks equivariance because the estimated coefficients are changed in a nonlinear way when any variable is rescaled (e. g. Malinvaud 1980, chap. 1).

3. HIGH-BREAKDOWN REGRESSION ESTIMATORS

This paper deals with linear regression methods that have the desirable property of equivariance. With respect to these estimators, an important robustness theorem states that "*50% is the highest possible value for the*

breakdown point, since for larger amounts of contamination it becomes impossible to distinguish between the good and the bad parts of the sample" (Rousseeuw and Leroy 1987, p. 120, italics added). Actually, 50% is an asymptotic value for the maximum breakdown point; in finite samples, the value is reduced by a degrees-of-freedom adjustment.

Which linear regression estimators attain the maximum breakdown point ? Like the univariate average, least squares regression is quite vulnerable to aberrant observations not only in the dependent variable ("regression outliers") but also among the regressors ("bad leverage points"). Either sort of data problem can produce a large bias, so least squares has a breakdown point of 1/n. On the other hand, the L_1 norm minimizes the sum of the absolute values of the residuals; as such, it estimates the conditional median of the dependent variable and might be expected to inherit the robustness of the univariate median. The L_1 norm is in fact highly resistant to regression outliers, but it performs no better than least squares when there are bad leverage points among the regressors; so its breakdown point is also 1/n (Rousseeuw and Leroy 1987, chap. 1 and 3).

The maximum breakdown point is attained by least median of squares, least trimmed squares, S-estimators, and other procedures that behave like multivariate versions of the mode. These estimators are based on the 50% of observations that cluster most tightly around the regression plane, and they are unaffected by data lying outside that cluster. As a result, the high-breakdown methods are quite robust, but they are inefficient when the data set is

uncontaminated. A researcher can apply a high-breakdown estimator initially and follow up with a more efficient estimator once any anomalous observations have been identified, scrutinized, and either reinstated, down weighted, or removed (Rousseeuw and Leroy 1987, chap. 1 and 3; Yohai 1987; Yohai and Zamar 1998).

4. THE CONTINUITY ASSUMPTION

To prove the theorem for the maximum breakdown point, Rousseeuw and Leroy (1987, p. 125) assume that all the model's regressors are continuous-valued ("in general position"). Since dummy variables obviously violate that assumption, the theorem must be modified before it can be applied to the models with which this paper is concerned. For some insight on the key role of the continuity assumption, it is helpful to examine an algorithm often used to compute high-breakdown linear-regression estimates. It must be emphasized that this discussion is heuristic since the theorem on the maximum breakdown point is independent of any particular algorithm or estimation strategy.

For the robust estimation of p linear regression coefficients, the elemental-set algorithm selects at random and without replacement p observations from the sample of n data. This elemental set is just sufficient to "estimate" the p regression coefficients, which in turn generate n residuals. A robust criterion is then applied to the residuals; for example, the median squared residual is computed for least median of squares regression. If n is small, this procedure can be repeated for all $n! / p!(n-p)!$ elemental sets, after which the coefficients with the smallest median squared residual are considered to be the high-

breakdown estimates. For larger n, it is computationally impractical or infeasible to examine every elemental set. It is also unnecessary, since the evaluation of a few thousand sets can be shown to locate high-breakdown estimates with near certainty (Rousseeuw and Leroy 1987, chap. 5).

The algorithm assumes that there is no linear dependency in an elemental set; for if the p observations are collinear, they cannot produce usable regression coefficients. But dependency is essentially precluded if all the regressors are continuous valued; a singular set of p equations could occur only as a fluke. In other words, the continuity assumption guarantees that all the sample data are available for inspection and evaluation.

Although dummy variables are designed to have full rank with respect to the entire sample, they will be linearly dependent in some sets containing at least p observations. In a model with many dummy variables, a lot of sets will be useless for generating estimates of coefficients. Because dummy variables reduce the amount of available data, the estimator's breakdown point necessarily deteriorates. Mili and Coakley (1996, p. 2598) give a proof of this conclusion, showing how the best breakdown point varies inversely with the amount of linear dependency.

5. DUMMY VARIABLES AND THE BREAKDOWN POINT

With these concepts in hand, let us examine the linear model

$$y = \alpha + X\beta + D\delta + u \quad , \qquad (1)$$

where y is a vector of n observations on a continuous-valued dependent variable; X is a matrix of n observations on p continuous-valued regressors; D is an n x k

array of dummy variables; u is a vector of n Gaussian variables independently and identically distributed with zero expectation and variance σ^2 ; α is a scalar intercept parameter; β is a vector of p slope parameters; and δ is a vector of k coefficients for the dummies. While y, X and D are data, α, β, δ, σ and u are unobserved. X and D are assumed to have full column rank, and the expectations of X'u and D'u are assumed to be zero vectors. The object is to obtain reliable estimates of α, β, δ and σ (Greene 2003, chap. 7).

Contamination enters this conventional model if some elements of u are replaced by numbers whose absolute values are arbitrarily large, thereby producing regression outliers; in addition, some elements of X may be altered without any corresponding changes in y, leading to bad leverage points. Suppose that D contains just one dummy variable; for example, a bankrupt enterprise might be coded by 1 and a solvent enterprise by 0. Suppose moreover there are as many solvent firms as bankrupt firms. Then how much contamination can a high-breakdown regression estimator handle ? If more than one fourth of the data is corrupted and it all happens to be in the group of solvent firms, then no equivariant linear regression estimator can be guaranteed to avoid breakdown with respect to the coefficient of the dummy variable since more than one half of the solvent group's data is inconsistent with the model. Of course, the same conclusion applies if the contamination happens to be concentrated in the bankrupt firms.

An equal number of solvent and bankrupt firms is the best case from the standpoint of robustness. For example, if just 20% of the firms are bankrupt, then

the contaminated observations in that group must be less than 10% of the entire sample to guarantee that breakdown can be avoided.

Consider now a data set for a cross-section of households. One of the variables shows whether annual household income averages less than $15,000, or between $15,000 and $50,000, or more than $50,000. To represent these categories, a researcher can put three dummies into D, suppressing the intercept α; or α can be retained if one of the dummies is omitted. However, the choice of parameterization has no effect on the breakdown point: the most robust situation still corresponds to an equal number of households in each income category, in which case the contaminated observations in any category must not exceed one sixth of the whole sample if breakdown is to be precluded. As before, the actual breakdown point can be lower than one sixth if most of the contamination happens to appear in a single group, an outcome that may be more likely if some group contains less than one third of the observations.

The preceding examples point to three conclusions about model (1). First, *the best-case breakdown point depends primarily on the number of groups in the most complex categorical variable*. The solvency/bankruptcy variable has just two groups, and the maximum breakdown point is about one fourth; the income variable has three groups, and maximum breakdown point is about one sixth. If the most complex categorical variable in a data set has k groups, then the best-case breakdown point is approximately 1/2k, as Hubert (1997) shows. Second, the actual breakdown point can be less than 1/2k, especially if some groups contain a small number of observations. And third, if contamination spoils the

estimate of a particular dummy-variable coefficient, the estimates of other coefficients are also likely to be biased simply because all the coefficients in a multiple linear regression are estimated jointly. (It is true that the dummies for a particular categorical variable are constructed to be orthogonal, but estimates of coefficients for different categorical variables may well be correlated.)

6. ROBUSTNESS IN MODELS WITH FIXED OR RANDOM EFFECTS

It is worthwhile to consider the implications of the previous section for robust estimation of the fixed-effects models popular in economics and other social sciences (Greene 2003, chap. 13). If a panel data set contains several years of observations on each of the 50 states of the U. S. A., a researcher might want to estimate, among other parameters, a fixed effect for each state. For computational reasons, these effects may be represented as group means rather than as explicit dummy variables; but again the choice of parameterization has no bearing on the breakdown point, which is at best 1/2k = 1/100. So breakdown cannot be ruled out if the anomalous observations for any state exceed 1% of the whole sample. For many types of economic data, that level of contamination does not seem unlikely.

While this author is not aware of robustness theorems that deal specifically with the random-effects model, the situation appears similar to the fixed-effects case. In terms of equation (1), Dδ disappears; and the estimation of β depends on the prior estimation of variance components in groups of residuals. There exist several high-breakdown estimators of dispersion (e. g. Rousseeuw and Croux 1993), but it seems that they may also fail if the contamination in any

group exceeds the bound 1/2k mentioned in the previous section. The random-effects model is routinely applied to panels containing rather short time series on several thousand individuals or households. If a variance component is required for every such unit, it is hard to see how the robustness of the estimated regression coefficients can be guaranteed.

Since there is no obvious remedy for the fragility of these estimators, researchers may want to be cautious about using models that partition data sets into many small groups. To the extent possible, researchers could replace dummies with continuous-valued regressors, especially where the goal is to control for heterogeneity. In a sample of the 50 U. S. states, for instance, it would be worth considering whether continuous-valued regressors measuring population, income and territorial extent could substitute for fixed or random effects. If so, the application of a high-breakdown linear regression estimator, followed by an efficiency improvement of the type mentioned above, would be an important element in any strategy designed to produce reliable parameter estimates.

REFERENCES

Croux, C., and Rousseeuw, P. J. (1993). "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association* 88, 1273-1283.

Donoho, D. L., and Huber, P. J. (1983), "The Notion of a Breakdown Point," in *A Festschrift for Erich Lehmann*, eds. P. Bickel, K. Docksum and J. L. Hodges, Belmont, CA: Wadsworth, 157-184.

Greene, W. H. (2003). *Econometric Analysis* (5th ed.), Upper Saddle River, NJ: Prentice Hall.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions,* New York: Wiley.

Hubert, M. (1997), "The Breakdown Value of the L_1 Estimator in Contingency Tables," *Statistics & Probability Letters* 33, 419-425.

Hubert, M., and Rousseeuw, P. J. (1997), "Robust Regression With Both Continuous and Binary Regressors," *Journal of Statistical Planning and Inference* 57, 153-163.

Malinvaud, E. (1980), *Statistical Methods of Econometrics,* Amsterdam: North-Holland.

Maronna, R. A., and Yohai, V. J. (2000), "Robust Regression With Both Continuous and Categorical Predictors," *Journal of Statistical Planning and Inference* 89, 197-214.

Mili, L., and Coakley, C. W. (1996), "Robust Estimation in Structured Linear Regression," *Annals of Statistics* 24, 2593-2607.

Rousseeuw, P. J., and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley.

Yohai, V. J. (1987), "High Breakdown-point and High Efficiency Estimates for Regression," *Annals of Statistics* 15, 642-665.

Yohai, V. J., and Zamar, R. H. (1998), "Optimal Locally Robust M-estimates of Regression," *Journal of Statistical Planning and Inference* 64, 309-323.

Zaman, A., Rousseeuw, P. J., and Orhan, M. (2001), "Econometric Applications of High-breakdown Robust Regression Techniques," *Economics Letters* 71, 1-8.