Sensitivity Bounds for use with Flawed Data

Denzil G. Fiebig and Pierre-Francois Uldry

Department of Econometrics and Department of Marketing, University of Sydney

Abstract: Data do not behave and applied work is made difficult by the need to cope with the myriad of problems that arise from flawed data. Prominent examples are data that contain missing values or where variables are only available in categorised form. Standard solutions include omitting incomplete records and replacing missing or categorised observations by some representative values. In these, and other cases of flawed data, there is some information available on the potential range into which any particular observation could feasibly fall. We contend that there is considerable diagnostic value in exploiting this information to compute bounds for the coefficient estimates and related statistics such as *t*-ratios. While one of the standard solutions may ultimately be used to produce a set of estimates, the bounds provide an indication of how sensitive these results are to the particular solution chosen. This approach is developed and illustrated by way of several examples.

"If the data were perfect, collected from well designed randomised experiments, there would be hardly room for a separate field of econometrics." Griliches (1986, p. 1466)

1. INTRODUCTION

Data do not behave and applied researchers are constantly faced with the problem of having to investigate substantive issues with flawed data. Problems such as missing data or data that are available only in grouped form are often "solved" by substituting some representative value. With grouped data the midpoint of the group interval may be used. Estimation procedures then proceed ignoring the initial problem.

In circumstances where there exists some information regarding the range of possible values associated with particular flawed data values, we propose the use of bounds as a diagnostic to convey the degree of uncertainty caused by the flawed data. For example, consider estimation of the mean of a random variable where the available data are in grouped form. The individual y_i are not observed but rather we know that y_i is contained in the interval (y_i^L, y_i^U). In terms of obtaining an estimate for the mean of y, a natural approach is to find the mean of the interval midpoints. How sensitive is this estimate to the choice of representative values?

It is possible to bound the actual sample mean that could be obtained with the particular data set had the actual y_i's been available. It is our contention that this type of diagnostic information is useful and should be reported more often in empirical work. It illustrates the sensitivity of results; here the sample mean, to one of the assumptions made in the data analysis; namely the use of interval midpoints as representative values.

We illustrate how these bounds can be calculated for coefficient estimates in multiple regression and for other statistics such as *t*-ratios. In doing so, we extend the existing work of Cameron (1987), Farebrother (1994), Fiebig (1993) and Kooreman (1993). Examples that involve reworking published papers are provided to illustrate the approach and its potential.

2. BOUNDS AS DIAGNOSTICS

Consider a partitioned regression model:

$$(1) \quad y = Z\gamma + x\beta + u ,$$

where y, x and u are nx1 vectors, Z is an nxk matrix, γ is a kx1 coefficient vector and β is a scalar coefficient. Suppose that the individual observations on y and x are not observed. Instead, only limiting values for each data point are known. In other words, we know that $y_i^L \le y_i \le y_i^U$ and $x_i^L \le x_i \le x_i^U$; i = 1, ... , n. For the sake of exposition only one independent variable is categorised; the procedures to be developed readily extend to situations where two or more of the independent variables have this feature.

As has been discussed such situations may arise if data are grouped; either because the collection agency only asked for information in this form or if the data are available after grouping of individual responses. There are many other instances where such situations arise. Data may be subject to rounding error. Records may be scrambled to preserve confidentiality (see for example Strachan et al., 1997 and references therein) or because matching of records from different sources

may be incomplete (see for example Kalb, 1997). Yet another situation is the classical case of missing data where there are known, or at least predictable, limits on the possible responses.

Ultimately, complete analysis of data of this type will involve some "solution"; midpoints may be substituted for grouped data or missing data may be imputed or simply ignored. In the case of a multiple linear regression model where the dependent variable is grouped the procedure of using group midpoints can lead to substantial biases in the estimated slope coefficients if the true underlying distribution is skewed; see for example Cameron (1987). Choice of an appropriate distribution to enable maximum likelihood estimation may be problematic.

Alternatively, it is possible to bound the coefficient estimates in this case and these bounds enable the researcher to know whether the choice of strong parametric assumptions are important or not. Thus, irrespective of the solution used, bounds can provide useful diagnostic information. They reflect what can be learnt from the data alone, without imposing any additional structure on the problem.

Coefficient bounds are obtained by solving well-defined optimisation problems. The upper bound on b, the OLS estimator of β, is the solution to the following problem:

$$(2)\quad Max \quad b = (x'M_Z x)^{-1}(x'M_Z y)$$
$$subject\ to\ L_x \le x \le U_x;\ L_y \le y \le U_y,$$

where $M_Z = I - Z(Z'Z)^{-1}Z'$, $L_x = (x_1^L, \cdots, x_n^L)'$, $U_x = (x_1^U, \cdots, x_n^U)'$, $L_y = (y_1^L, \cdots, y_n^L)'$ and $U_y = (y_1^U, \cdots, y_n^U)'$.

If only the dependent variable is categorised, the problem admits an analytical solution; see Fiebig (1993) and Farebrother (1994). When it is the independent variable that is categorised, one has to solve a somewhat more difficult non-linear optimisation problem that does not admit an analytical solution. However, Kooreman (1993) provides the Kuhn-Tucker conditions that characterise the solution. Combining these situations allows both independent and dependent variables to be categorised. Uldry (1996) has obtained the Kuhn-Tucker conditions that characterise the solution to this extended problem.

Another extension to the existing work is to produce bounds on various regression statistics as well as the coefficient estimates. Probably the most useful is to provide maximum and minimum t-statistics which would indicate the sensitivity of inferences. For the maximum t-statistic the optimisation problem is:

$$(3)\quad Max \quad t_b = \frac{b}{\sqrt{s^2(x'M_Z x)^{-1}}}$$
$$subject\ to\ L_x \le x \le U_x;\ L_y \le y \le U_y,$$

where s^2 is the estimate of the disturbance variance.

An important motivation for this extension is that extreme values for the t-statistics *do not* necessarily correspond to extreme values for the coefficient estimates.

Consider the following simple example adapted from Grimmett and Ridenhour (1996). A sample of 5 observations is given by {0, 1, 1, 2, y_5} where it is known that $0 \le y_5 \le 2$. Using only the 4 "good" observations the estimate of the mean of y is 1. Bounds on the estimate are obtained by replacing y_5 by either 0 or 2, yielding extreme estimates for the sample mean of 0.8 and 1.2. Associated with these extreme coefficient estimates are t-statistics of 2.138 and 3.207. However, the maximum t-statistic for admissible y values is 3.317 which occurs when $y_5 = 1.5$.

Notice that the bounds on the coefficient estimates are not confidence intervals. They do not rely on the notion of repeated samples, rather they represent the bounds on the numerical values of estimates for the particular data set being analysed.

Other than the case of coefficient bounds when only the dependent variable is categorised, the calculation of bounds involves optimisation problems that need to be solved numerically. In the examples to be presented, the SOLVER of EXCEL has been used successfully to produce the required results.

3. APPLICATIONS

3.1 Byron and Ashenfelter (1995)

Following previous attempts to predict wine prices by econometric methods, Byron and Ashenfelter (1995) estimated regression models that attempted to explain the price paid at auction for various vintages of Australia's Grange Hermitage. They estimated models of the form:

$$(4)\quad \log(price) = \beta_0 + \beta_1 age + \beta_2 rain + \beta_3 temp + \beta_4 temp^2 + \beta_5 diff + u$$

where *price* is an average auction price calculated over data from either 1991, 1992, 1993 or an average of these three log prices; *age* is the age of the wine; *rain* is rainfall, *temp* is temperature and *diff* is a measure of temperature extremes. See Byron and Ashenfelter (1995) for more complete descriptions and the actual data.

Over the range of vintages from 1959 to 1987 the explanatory variables were complete but some vintages were not traded in some years leading to missing values in the dependent variables. Table 1 provides the pattern of missing values for the three vintages that are affected.

Table 1: Pattern of missing values*

Vintage	*price91*	*price92*	*price93*
1960	-	210.50	277.67
1986	-	91.00	100.08
1987	-	-	65.00

* Source: Byron and Ashenfelter (1995).

In some initial testing, separate equations were estimated with dummy variables included for the vintages containing missing observations. On the basis of these tests it was decided to proceed with an equation where the dependent variable was defined as the composite price given by the average of the three log prices. This regression used a reduced sample from 1961 to 1985 formed by deleting the incomplete observations as well as the 1959 observation.

While there are no known restrictions on the missing average prices, we can provide some sensible limits on the basis of the data that are available. These values have been provided in Table 2. These limits can then be translated into limits on the composite dependent variable.

Table 2: Limits defined for missing values

Vintage	$price_{91}$		$price_{92}$	
	Lower	Upper	Lower	Upper
1960	100	280	(210.5)	(210.5)
1986	50	101	(91)	(91)
1987	40	65	45	75

Note: Numbers in parenthesis are not missing.

With these limits it is now possible to compute bounds that reflect the sensitivity of the regression results to the deletion of incomplete observations. Table 3 presents the bounds on the OLS coefficient estimates together with the coefficient estimates obtained from a regression using the reduced sample. Although similar, the latter do not correspond exactly to the regression results reported by Byron and Ashenfelter (1995). Despite some effort, we can not be sure of the cause of these differences and hence have proceeded on the basis of the reported data and the description of the modelling procedure given by Byron and Ashenfelter (1995).

Bounds on the coefficient estimates and those calculated from the reduced sample always have the same sign. No matter what the values taken by the missing data, within the prescribed limits, there is no ambiguity in the estimated signs of the marginal responses. However, there is some sensitivity reflected by the observation that for four of the five variables, the reduced sample estimate does not fall within the bounds. Potentially, the omitted observations are not consistent with the relationship estimated on the basis of the complete data.

Table 3: Coefficient estimates and their bounds

Variable	*Reduced sample*	*Bounds*	
		Minimum	Maximum
Intercept	-57.5953	-54.9981	-26.4689
Age	0.0428	0.0389	0.0454
Rain	-0.0034	-0.0047	-0.0041
Temp	6.6064	3.4185	6.4983
$Temp^2$	-0.1732	-0.1717	-0.0902
Diff	-0.2027	-0.3451	-0.2706

Table 4 has the same format as Table 3 but presents the *t*-statistics. The sensitivity observed in the coefficient estimates is also reflected in the *t*-statistics. Notably, the bounds suggest that the inferences associated with the estimated temperature responses are sensitive to the treatment of the missing data. While the reduced sample results indicate both temperature coefficients are reasonably precisely estimated, the bounds suggest that this picture may be overly optimistic. Both of the *t*-statistics have maximal magnitudes less than the reduced sample results and the range of possible values includes situations where estimated responses would not be significantly different from zero.

Table 4: *t*-statistics and their bounds

Variable	*Reduced sample*	*Bounds*	
		Minimum	Maximum
Intercept	-2.58	-0.96	-1.99
Age	9.48	7.83	10.79
Rain	-3.28	-3.02	-3.82
Temp	2.76	1.16	2.19
$Temp^2$	-2.71	-1.14	-2.16
Diff	-2.07	-2.18	-3.13

Note: The minimum and maximum bounds refer to the magnitude of the *t*-statistics.

On the other hand, the impact of the age of the wine on the auction price is not sensitive to the treatment of the incomplete observations. Out of all of the estimated coefficients in the reduced sample, only that associated with *age* fell within the bounds. Moreover, this impact was always precisely estimated, irrespective of the values taken by the missing price data.

3.2 Bloch (1992)

Bloch (1992) was interested in the impact of the rate of change of the price of competing foreign products and the rate of change of costs of domestic production on the rate of domestic price inflation. Empirical evidence was provided by OLS estimation of an equation of the

form:

$$(5) \quad \dot{p}_d = [\gamma_1 + \gamma_2 CR]\dot{c}_d + [\gamma_3 + \gamma_4 CR]\dot{p}_f + u,$$

where $\dot{p}_d$ is the average rate of change in the price of domestic product, CR is industry concentration, $\dot{c}_d$ is the average rate of change in the marginal cost for the domestic product and $\dot{p}_f$ is the average rate of change in the price of competing foreign product. See Bloch (1992) for more complete descriptions and a data listing.

Using the published data it is not possible to exactly reproduce the regression results reported in Bloch's paper. The culprit is data rounding. Bloch has generated the variables used in the regression analysis from another set of variables but has only reported rounded values of the generated data. Consequently, all data points are subject to error but we know the limits of this error.

For this example, our bounds analysis is used to:

a) Check that the reported results fall within our calculated bounds in order to complete the replication exercise;
b) Indicate the diversity of results that can potentially occur because of small, albeit pervasive, perturbations in the data.

Bloch's OLS coefficient estimates for the parameters of equation (5) are reproduced in Table 5 together with their associated bounds. For this example, the restrictions imposed on the data refer to the 4 distinct variables rather than the actual variables used in the regression. This avoids inconsistencies that would arise if variable limits were imposed directly on the interaction variables. Results for the *t*-statistics are presented in Table 6.

In terms of our first objective, the replication exercise has been successful. All of the reported results, both estimated coefficients and *t*-statistics, fall within the bounds. Researchers intending to work further with these data can do so with reasonable confidence. Naturally, obtaining the original data remains preferable.

This example illustrates the potential sensitivity of regression results to rounding error. In particular, consider the estimates for the rate of change in the price of competing foreign product. Using the reported data that is subject to rounding error, one can not exclude the possibility of a precisely estimated negative response at one extreme, or, a positive but imprecisely estimated response at the other.

Because we have the advantage of knowing Bloch's reported results we know that the actual data support a negative but imprecisely estimated impact for the rate of change in the price of competing foreign product. However, what if a prospective researcher is interested in extending the analysis by considering a modified specification not considered by Bloch? For this particular exercise, our type of bounds analysis would be useful in highlighting the potential sensitivity of estimates and their associated inferences caused by the uncertainty associated with the reported data. More generally it serves to illustrate the potential fragility of the results of data analyses.

Table 5: Coefficient estimates and their bounds

Variable	*Reported*	*Bounds*	
		Minimum	Maximum
Intercept	0.0081	-0.0184	0.0345
$\dot{c}_d$	1.0961	0.9599	1.2262
$\dot{p}_f$	-0.1451	-0.4449	0.1622
CR. $\dot{c}_d$	-0.6103	-0.8491	-0.3662
CR. $\dot{p}_f$	0.5061	0.2773	0.7247

Note: The bounds use the data provided by Bloch (1992), while the reported results use the original data that are not subject to rounding error.

Table 6: *t*-statistics and their bounds

Variable	*Reported*	*Bounds*	
		Minimum	Maximum
Intercept	0.91	-2.09	3.94
$\dot{c}_d$	21.19	13.25	36.04
$\dot{p}_f$	-1.47	1.61	-4.91
CR. $\dot{c}_d$	-5.79	-2.84	-10.80
CR. $\dot{p}_f$	5.28	2.43	10.02

Notes: (i) Minimum and maximum bounds refer to the magnitude of the *t*-statistics.
(ii) The bounds use the data provided by Bloch (1992), while the reported results use the original data that are not subject to rounding error.

4. CONCLUSION

The underlying philosophy for the diagnostics discussed in this paper is that empirical results are sensitive to choices that are routinely made in modelling and that some attempt should be made to convey the extent of this sensitivity. For the types of flawed data discussed in this paper, bounds can provide such information. They reflect what can be learnt from the data, without imposing any additional structure on the problem. Moreover, the diagnostics are easily calculated. All of the computations reported were generated using the spreadsheet program EXCEL.

While the examples that have been discussed provide bounds for OLS results, the underlying philosophy and approach are more widely applicable. In a non-regression context, there exists a considerable literature on the problem of bounding the moments of a univariate distribution from grouped data. See for example Gastwirth and Krieger (1975) and Krieger and

Gastwirth (1984). Here the emphasis has been on sharpening the bounds by incorporating information in addition to the interval widths. Such extensions would be useful in the regression context.

The Byron and Ashenfelter example suggests one possible extension to more complex estimation procedures. Typically in missing data problems, different estimation procedures are required depending on whether data are missing at random or not. In the newer terminology, if the pattern of missing data is ignorable then the parameters of interest can be consistently estimated from the complete data. The incentive to search for improved estimation procedures derives from the possibility of increased efficiency from using the extra but incomplete data. Alternatively, if the data are missing because of some sort of self-selection then there is the risk of biases as well as efficiency loss if the problem is ignored. If it is possible to assign limits to the missing observations, then bounds could be used to determine the sensitivity of coefficient estimates and hence to indicate the need for the estimation of a selection model.

5. ACKNOWLEDGEMENTS

We would like to thank Ray Byron for his assistance. Financial support from the Australian Research Council is also gratefully acknowledged.

6. REFERENCES

Bloch, H., Pricing in Australian manufacturing, *The Economic Record*, 68, 365-376, 1992.

Byron, R.P. and Ashenfelter, O., Predicting the quality of an unborn Grange, *The Economic Record*, 71, 40-53, 1995.

Cameron, T.A., The impact of grouping coarseness in alternative grouped data regression models, *Journal of Econometrics*, 35, 37-58, 1987.

Farebrother, R.W., Bounds on coefficient estimates when the dependent variable is grouped, Solution 93.1.1 *Econometric Theory*, 10, 226, 1994.

Fiebig, D.G., Bounds on coefficient estimates when the dependent variable is grouped: A problem, *Econometric Theory*, 9, 145, 1993.

Gastwirth, J.L. and Krieger, A.M., On bounding moments from grouped data, *Journal of the American Statistical Association*, 70, 468-471, 1975.

Griliches, Z., Economic data issues, Ch. 25 of Z. Griliches and M.D. Intriligator eds. *Handbook of Econometrics*, North Holland, 1986.

Grimmett, D.R. and Ridenhour, J.R., The effect of a variable data point on hypothesis tests for means, *The American Statistician*, 50, 145-150, 1996.

Kalb, G. The effect of unemployment benefits on labour supply, *Proceedings of the Econometric Society Australasian Meeting: Volume 4*, 833-871, 1997.

Kooreman, P., Bounds on the regression coefficients when a covariate is categorised, *Communications in Statistics - Theory and Methods*, 22, 2373-2380, 1993.

Krieger, A.M. and Gastwirth, J.L., Interpolation from grouped data for unimodal densities, *Econometrica*, 52, 419-426, 1984.

Strachan, R., King, M.L. and Singh, S., Regression analysis using scrambled or hidden responses, *Proceedings of the Econometric Society Australasian Meeting: Volume 2*, 89-106, 1997.

Uldry, P-F., A bounds approach to missing data, Master of Economics Essay, University of Sydney, 1996.