MEASUREMENT ERRORS IN INTRODUCTORY ECONOMETRIC COURSES

John H. Herbert
Virginia Polytechnic Institute and State University
and
Energy Information Administration

INTRODUCTION

This article describes an instructive method for presenting the implications of measurement error in a regression analysis for the introductory econometrics course. As distinct from other methods, this method can be used to obtain estimates of the linear relationship between measured variables without requiring specific information on the relative magnitude of the measurement errors. However, the chosen method yields a set or range of estimates for a regression coefficient rather than a point estimate. This range, commonly referred to as bounds for a regression coefficient, indicates the uncertainty in a regression coefficient due to measurement error in one or more explanatory variables.

The estimated range is not to be confused with a confidence interval, which indicates the uncertainty in a regression coefficient due to sampling error. In applications the bounds may be found to yield a very large range of estimates unless there is a strong linear relationship between the variables included in the regression equation.[1] This result may be troubling to some, yet it has instructive value for students. It draws attention to the importance of obtaining information on the measurement error in measured variables used in empirical analysis. If such information is available, several possibilities emerge, depending on the type and quality of the information collected on the measurement error. The analyst might be able to use a consistent point estimator, in which case the analyst could rest assured that a value estimated from data would approach the value for a population (the parameter) as the sample size was increased [Fuller, 1987]. Second, the analyst could conduct a sensitivity analysis to evaluate the bias in the conventional point estimate, that is, the difference between the expected value of the statistic estimated from the data and the population parameter [Herbert and Kott, 1988].

Since it has been found that some students respond well when a method is presented in the context of its development, the material is presented in this way. This has additional advantages for an introductory course. The students acquire some sense of the development of econometrics as a discipline. Moreover, ideas rather than models are emphasized since earlier treatments of the method relied less on assumptions and more on intuition. For example, the estimated lines from which the bounds are obtained are considered as a consequence of viewing the data from different perspectives. First, we plot a variable y against another variable x and then fit a line from the perspective of y. This gives us one bound. Then we plot the same variable x against the same y and then fit another line from the perspective of x. This gives us the other bound. Thus, the bounds are a consequence of viewing the data from different perspectives.

BACKGROUND

The approach presented here was developed by Ragnar Frisch in the late 1920s and the early 1930s. Frisch coined the term econometrics and was the founder of *Econometrica*, the first journal devoted to the use of statistical and mathematical methods to enhance economic analysis. Frisch, who shared the first Nobel prize in economics with Jan Tinbergen, considered random measurement error rather than random equation error to be the major statistical problem in estimating economic behavioral relationships.

In *Statistical Confluence Analysis*, Frisch [1934] identified his prescriptions for and opinions about a research program for applied econometric research. It summarized and extended much of his thinking about statistical economics of the previous ten years. A principal tool delineated and applied in the text is the estimation of bounds for a regression coefficient in a linear relationship between economic variables. The bounds were recommended by Frisch for general use since he expected that most explanatory variables in a regression analysis with economic variables would contain a measurement error.

The bounds were developed in a non-parametric setting with minimal assumptions about the distribution either of the measurement errors or the variables. This is consistent with Frisch's overall approach to statistical analysis in which assumptions about variables were kept to a minimum in the development of a procedure because he believed that in many instances such information, such as whether the errors in the variables were normally distributed, would not be known. Thus, the interpretation of estimated regression coefficients and related summary statistics should not be based on the satisfaction of such assumptions.

Our treatment of Frisch's technique is largely graphical and non-technical. The objective is to present the problem of measurement error in such a way that the majority of students attain an understanding of the general nature of the problem.

MEASUREMENT ERROR

A student will often find in many empirical investigations that only proxy or poorly measured variables are available for estimating relationships between variables. These proxy or poorly measured variables differ in magnitude from a desired variable; that is,

$$(1) \qquad x_t \neq \boldsymbol{X}_t \,,$$

$\boldsymbol{X}_t$ = desired measurements on variable $\boldsymbol{X}$ for t $(t = 1, \dots, T)$ observations;
x_t = available measurements on variable $\boldsymbol{X}$ for t observations.

The reason why x_t is not equal to $\boldsymbol{X}_t$ is because of a measurement error (u_t). Thus, the observed x_t can be expressed as being equal to the desired measurement plus a measurement error; that is,

$$(2) \qquad x_t = \boldsymbol{X}_t + u_t.$$

If the u_t were known we could obtain $\boldsymbol{X}_t$ from x_t. Since the u_t are not known we need to make some assumptions about them.

It is assumed that the observed measurement varies from the desired measurement in a nonsystematic or random way.[2] In some instances the available measurement is above the desired measurement, in which case the measurement error in equation (2) can be viewed as having a positive value. In other instances the available measurement is below the desired measurement, in which case the measurement error can be viewed as having a negative value. In addition, these positive and negative values do not follow a pattern. Positive errors do not consistently follow positive errors and negative errors do not consistently follow negative errors nor do positive errors consistently follow negative errors. Positive and negative errors tend to balance or cancel one another. Thus, the u_t have an expected value of zero for each observation.

It is also assumed that there is a linear relationship between two variables y and $\boldsymbol{X}$ summarized by a coefficient B; that is,

$$(3) \qquad y_t = \mathrm{a} + \mathrm{B}\boldsymbol{X}_t + e_t.$$

The e_t are also random errors with an expected value of zero for each observation. These errors could be a consequence of variables left out of the equation that are useful for explaining y_t or they could be a consequence of the irreproducibility of human behavior. They could also be due to measurement error in y_t. If we could control for all omitted variables and all measurement errors in an empirical analysis, and if the irreproducibility of human behavior was not viewed as a problem, then we would get the exact value for B from the data; that is,

$$(4) \qquad \boldsymbol{Y}_t = \mathrm{a} + \mathrm{B}\boldsymbol{X}_t.$$

Thus, the observed variable y_t can be expressed as,

$$(5) \qquad y_t = \boldsymbol{Y}_t + e_t.$$

The e_t and u_t are not expected to be related to each other or to $\boldsymbol{X}_t$ or $\boldsymbol{Y}_t$; that is, the e_t and the u_t are independent of each other and of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. However, the assumption of independence is not always realistic and more advanced treatments of the bounds relax the independence assumptions [Bekker, et. al., 1987; Erickson, 1993; Krasker and Pratt, 1986; Kroch, 1988].

TABLE 1
Data

Y	*X*	e	u	y	x
3.0	3.0	0.0	0.0	3.0	3.0
4.0	4.0	2.0	1.0	6.0	5.0
5.0	5.0	-1.0	0.0	4.0	5.0
6.0	6.0	-1.0	0.5	5.0	6.5
7.0	7.0	0.0	-0.5	7.0	6.5

For most empirical analyses the trick is to obtain estimates for B given that *observed Y* (*y*) has a random error associated with it and *observed X* (*x*) has a random measurement error associated with it, and little or no information is available about the magnitudes or other properties of these errors. This is the major problem addressed by Frisch more than fifty years ago.

ESTIMATING LINEAR RELATIONSHIPS

When a random error is associated with an observed variable, a sensible procedure to follow in obtaining estimates for a slope B is to minimize the squared difference (which difference is due to the random error) between the observed value and the line. Now, let us set B equal to one and graph the line relating *Y* and *X* (see Figure 1).

The numbers in the first two columns of Table 1 are simply numbers representing the exact relationship between the variables. Next, consistent with (2) and (5), we randomly assign some error (*e* and *u*) to *Y* and *X* respectively, and designate these variables as *y* and *x* in columns 5 and 6 of Table 1. Thus, the *y* and *x* values in the last two columns of Table 1 are the sum of the true values reported in columns 1 and 2 and the random errors reported in columns 3 and 4. Then, we examine some situations that occur as a consequence of (2) and (5).

First, we examine the case where the random error that prevents us from obtaining B exactly is associated solely with *Y*. This is represented by the graph of *y* against *X* (see Figure 2). The line in Figure 2 is obtained by minimizing the squared difference between observed value of *y* and the line as indicated by the direction of the arrowhead. As a result of this ordinary least squares minimization we obtain an estimated value for B, the slope of the line relating *y* and *X*.

Second, we examine the case where all the random error that prevents us from obtaining B exactly is associated with *X*. This is represented by the graph of *Y* against *x* (see Figure 3). The line in this figure is obtained by minimizing the squared difference between the observed value *x* and the line. The direction of minimization is again indicated by the direction of the arrowhead.

FIGURE 1
The Relationship between Y and X

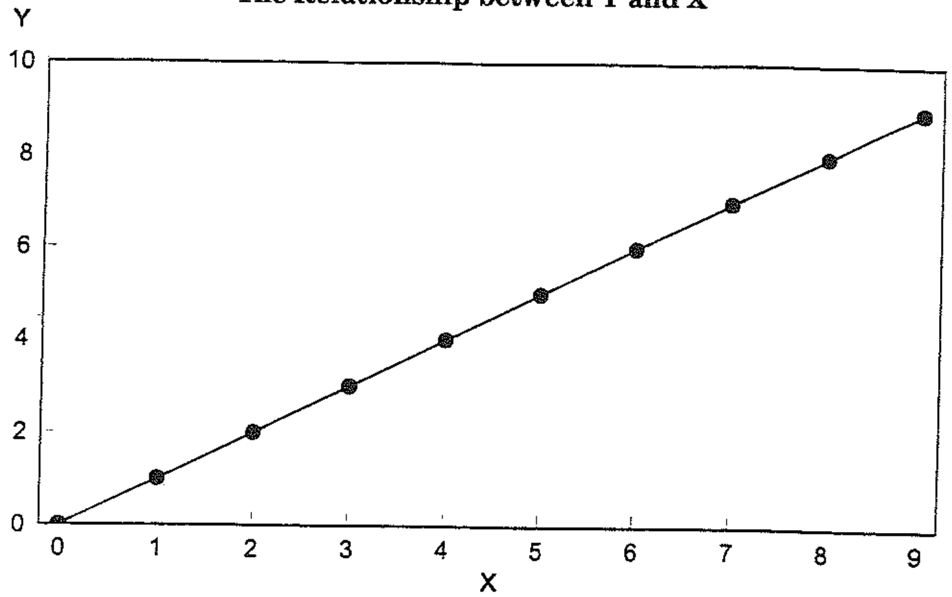


FIGURE 2
The Relationship between y and X

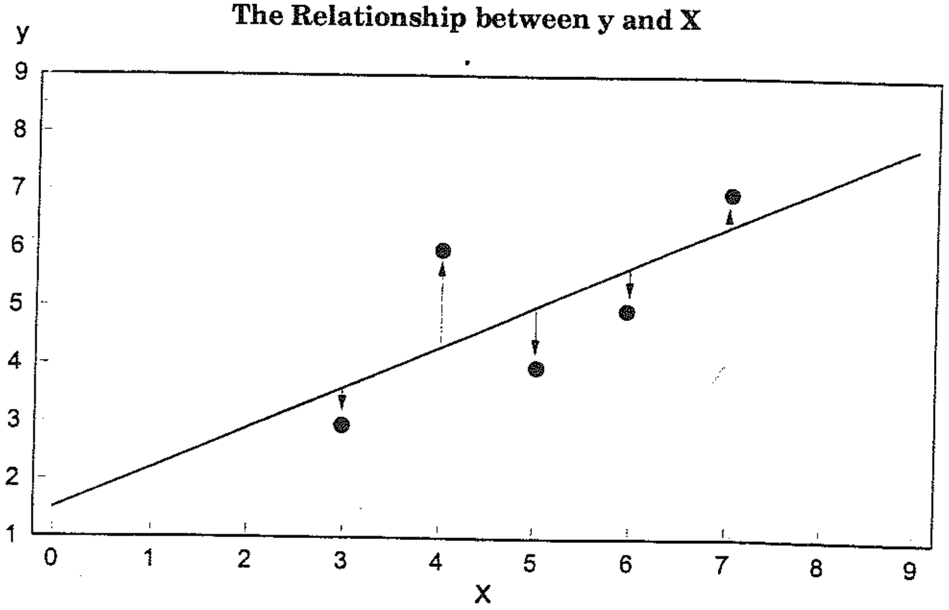


FIGURE 3
The Relationship between Y and x

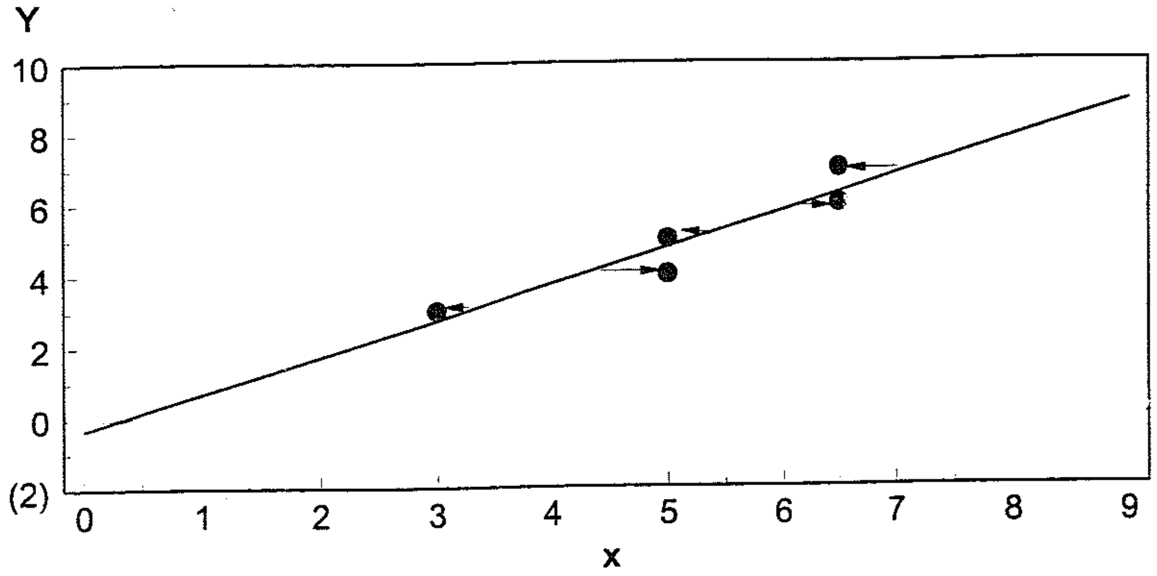
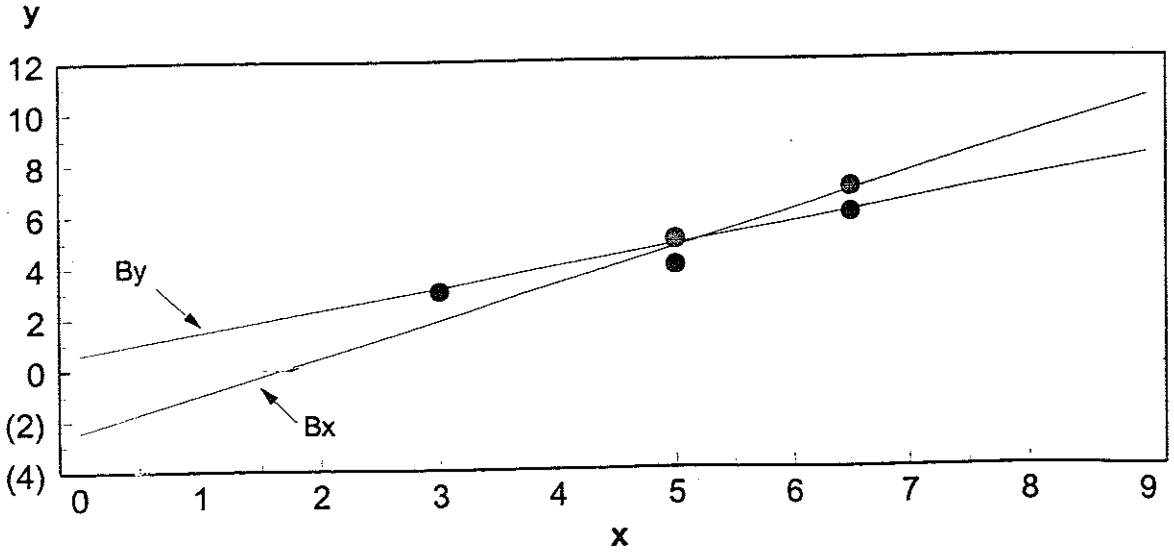


FIGURE 4
The Relationship between y and x



Finally, we examine the graph of y against x (see Figure 4). Now it is no longer easy to specify how the line representing the relationship between Y and X is to be estimated. In this situation Frisch recommended reporting bounds for the regression coefficient B. These bounds are obtained by first assuming that all the error is in the variable represented on the vertical axis; that is, we minimize in the direction of y. Second, we assume that all the error is associated with the variable represented along the horizontal axis; that is, we minimize in the direction of x.

The estimated coefficient when we minimize in the direction of y (B_Y) is +0.84; that is,

$$y = 0.614 + 0.84x.$$

The estimated coefficient when we minimize in the direction of x (B_X), and reexpress the equation with y on the left-hand side of the equals sign, is +1.43; that is,

$$x = 1.7 + 0.7y,$$

$$-0.7y = 1.7 + (-1.0)x,$$

$$y = 1.7/\text{-}0.7 + (-1.0/-0.7)x,$$

$$y = -2.43 + 1.43x.$$

Thus, the bounds are equal to the range +0.84 to +1.43. This range includes the coefficient 1, the exact value for B discussed previously.

In any application, no matter how many explanatory variables are measured with error, the student is taught to follow this same procedure; that is to minimize in the direction of each poorly measured variable and then to reexpress the estimated equations so that y, the natural dependent variable, is on the left-hand side of the equals sign. The range for a coefficient expressing the relationship between y and another variable is the bounds for the set of coefficients obtained in this way.

APPLICATION TO APPLIED PROBLEMS

Milton Friedman [Friedman, 1992; Friedman and Schwartz, 1991] has repeatedly emphasized that many estimation situations call for the reporting of not only the conventional least squares estimates but also the bounds. This is because of the extensive use of poorly measured or proxy variables in these investigations. A succinct example of such an application is presented next.

In order to plan deliveries of natural gas, utilities throughout the United States estimate the relationship between the use of natural gas in the households they serve and temperature [Trost, et al., 1994]. They are also interested in the relationship between gas use and price. Now, the use of natural gas is metered and well measured for most households and the price that customers pay for natural gas is also known.

However, exact temperature measurements for a household are not generally known. Only measurements at a particular site(s) where the households are located within the region, which may include several cities and the surrounding service area of a gas distribution company, are known. Information on the magnitude and other features of the measurement errors are also not available. Even though temperature has, by far, the greatest influence on gas use, variables such as income and household demographics also influence gas use. Yet, such excluded variables are expected to be independent of price and weather. The measurement errors associated with temperature are also expected to be independent of the equation error associated with gas use. Thus, without further information on the measurement errors, this description represents a problem for estimation in which the bounds along with the conventional point estimate are called for to provide a summary of the statistical relationship between gas use and temperature.

An equation is first estimated by minimizing in the direction of natural gas use per customer because of the equation error associated with it. This is the conventional least squares estimation with gas use as the dependent variable and temperature and price as explanatory variables. From this estimation we obtain the lower bound coefficient representing the relationship between gas use and temperature. An equation is next estimated by minimizing in the direction of temperature because of the measurement error associated with it. This estimation is then reexpressed with customer use of natural gas on the left-hand side and temperature and price on the right-hand side of the equals sign. This reexpressed equation provides us with the upper bound coefficient representing the relationship between gas use and temperature. For this application it is not necessary to minimize in the direction of the price of natural gas because the variable is well measured.

In a related application with real data some interesting results were obtained. In this application monthly, state-level values on gas use per customer, average price and other variables were used. All the variables were poorly measured. Now, the conventional least squares estimate for the own-price elasticity for this application was -0.560 with a standard error of 0.04. Instead of just reporting these estimates as statistics for summarizing the relationship between natural gas use per customer and price, the bounds were reported. It was found that the bounds for the own-price elasticity equaled -0.56 to -0.92 [Herbert, 1988]. Without using additional information on the measurement errors, such as the relative magnitude of the variance of the measurement errors to the variance of the variables, or without resorting to *ad hoc* assumptions, it was reported that the coefficient summarizing the relationship between gas use and price could lie anywhere between -0.56 and -0.92.

The fact that the conventional least squares estimate (-0.56) is 61 percent of the upper bound (in absolute value) estimate has practical significance. Estimates of own-price elasticities are used by gas distribution companies to estimate expected revenue streams. If the distribution company(ies) only used the conventional least squares estimate then the expected revenue from a price increase would possibly be greatly overestimated [Herbert, 1992].

In a much different analysis of aggregate data for the United States and the United Kingdom extending back to the 19th century, the consideration of bounds motivated Friedman and Schwartz [1982, 283] to modify an initial assessment on the stability of an estimated demand for money equation. Their demand for money equation included an income variable, proxy variables for the yield on money and for the yield on physical assets and several dummy variables. All coefficients were found to be highly significant with t-ratios greater than 3.5. Thus, it appeared that "money demand is stable over time —in the sense of a demand function— as well as between countries." The bounds for an income coefficient, representing an income elasticity, were consistent with this conclusion. The bounds for the income elasticity were 1.15 to 1.17. However, it was also discovered that the coefficients for the two proxy variables varied by a factor of more than 3 (-8.82, -27.82) for the yield on money proxy variable and by a factor of more than 4 (-.59, -2.59) for the yield on physical assets proxy variable.

In general, the discovery of a small range for a coefficient, such as that for the income elasticity, is additional evidence of the reliability of an estimated relationship. On the other hand, a large range for another coefficient points out the unreliability of the estimated relationship represented by the coefficient. In fact, the bounds for a coefficient may include zero when the conventional t-statistic indicates that the coefficient is different from zero.[3] Thus, the reporting of bounds frequently makes a difference in terms of an inference about a population parameter in many applications.

SOME ADDITIONAL HISTORY

Early applications of Frisch's methodology included a complete evaluation of the Cobb-Douglas production function by Horst Mendershausen [1937] and an evaluation of stock prices as part of a business cycle analysis by Haavelmo [1938]. The bounds methodology was also used extensively in the classic Stone [1954] demand studies as well as in Tinbergen's [1939] pathbreaking League of Nations study of the macroeconomy.

Koopmans in a classic doctoral dissertation published in 1937 evaluated in detail Frisch's method as well as alternative methods for performing a regression analysis when there are measurement errors in the variables. He also extended the theoretical aspects of the bounds methodology as did Reiersol [1945] in a very general sense.

One major weakness of the bounds approach pointed out by Koopmans [1937] is that it totally ignores the uncertainty in the estimated regression coefficients associated with the size of the sample as represented by the sampling or standard error of the estimated regression coefficient. This was a major issue early on because many data series were short. This is not a major problem if the sample is large since the standard error becomes smaller the larger the sample size. It is also not a problem if the data set is treated as representative of a population and not a sample.[4] Ruist [1947], moreover, in an infrequently referenced paper provided a methodology for obtaining sampling errors for the bounds and thus weakened Koopmans' major argument against the bounds methodology. Yet, even though he made no assumption about the magnitude of the variance of the measurement errors, it was still necessary for Ruist to assume that the variables were normally distributed.

Koopmans provided a point estimator in his book which allowed for both sampling error and measurement error. This estimator, however, also required additional assumptions. It required that the ratio of the variance of the measurement error in the one variable to the variance of the measurement error in the other variable be known and that the measurement errors be normally distributed.

Little attention was directed to the bounds methodology during the 1950s, 1960s and 1970s. Then Patefield [1981], Kalman [1982] (who continued with the nonparametric treatment of the bounds favored by Frisch), and Klepper and Leamer [1984] examined the Frisch bounds methodology with renewed enthusiasm. This later literature has spawned, in turn, other technical papers which address error terms that are not independent of each other and of variables [Bekker, et al., 1987; Erickson, 1993; Krasker and Pratt, 1986; Kroch, 1988].

CONCLUDING COMMENTS

Although the bounds methodology was presented here in a two variable context, the methodology was designed to be applied to equations with many variables. After the 1930s, the major emphasis for theoretical econometrics was on estimators for systems of behavioral equations. Theoretical progress in this area of study generally required restrictive distributional assumptions. Much applied work, however, involved the estimation of single linear equations in which it was unquestionably assumed that the error was in the left-hand side variable of the linear relationship.

The advantage for many applications of assuming that the right-hand side variable is error free was that a point estimate for the slope coefficient (B), as well as a standard error for this coefficient indicating its uncertainty, was obtained. Moreover, this standard error was similar to the well-known standard deviation for a mean in a univariate analysis. Thus, the statistical and economic interpretation of the relationship was facilitated by this choice.

As sample sizes available for an estimation increase, the bounds are a particularly useful indicator of the uncertainty in estimated coefficients. As the sample size increases, the standard errors associated with estimated regression coefficients necessarily decline. Hence, it is common that many regression coefficients are declared statistically significant since the magnitudes of most coefficients are several times as large as their estimated standard errors. In these evaluations of significance, the assignment of a unique value for a coefficient in a regression analysis is taken for granted. Yet, significant uncertainty about the regression coefficient remains because of the measurement error. This uncertainty is indicated by the bounds which do not necessarily decline as the sample size increases.

With the bounds it is necessary to consider a range of values for a coefficient. This consideration may greatly influence the economic interpretation of the results from the statistical analysis. Moreover, any uncertainty uncovered using the bounds technique can only be resolved by obtaining information about the measurement errors. Hence, some sample surveys are conducted today in which sample survey funds are not only allocated to maximize the size of the sample but to obtain information on the measurement error. Yet, until such surveys become commonplace and information is collected on the measurement errors in proxy variables, the bounds will remain a most useful tool for many econometric investigations.

NOTES

The author would like to thank Bob Trost of the economics department at George Washington University, Noel Uri of the Economic Research Service of the United States Department of Agriculture, Eric Slud of the math department at the University of Maryland, Phil Kott, senior mathematical statistician at the National Agricultural Statistical Service and Ted Biribin of SAIC for comments and other support. In particular, the author would also like to thank two anonymous referees and the editor of this *Journal* for providing the author with encouragement and several very useful and insightful comments.

1. Since cross-sectional data tend to yield estimated equations with weak linear relationships between the variables, and time series data tend to yield estimated equations with strong linear relationships between the variables, some might be tempted to state that the bounds methodology can be most usefully applied to time series data. However, any knowledge, no matter how modest, should guide the application of this methodology. If the measurement error in a poorly measured variable is considered too large to be ignored, the bounds methodology should be employed.
2. Systematic measurement error, where the measured variable is systematically above or below the correctly measured variable, is not much treated in econometrics or statistics despite the fact that this type of measurement error is relevant for some economic series. For a theoretical treatment of this type of measurement error see Herbert and Dinh [1989].
3. Such results were found in an interesting study by Klepper, Kamlet and Frank [1993] which included two case studies with real data. The paper also contains useful material on computing bounds as well as a description of a modern method for reducing the size of the bounds.
4. This would be the situation if the data were viewed as part of a particular case study rather than as a sample or as an experiment capable of replication. Then, the bounds are viewed as descriptive statistics necessary for a complete summary of the relationship between the variables in this population when there are errors in the variables. Knowledge is advanced by comparing the results of different case studies on similar material by the same investigator or by other investigators. General conclusions about relationships are based on the similarity of results across case studies. This was, in fact, the approach taken by Frisch in much of his empirical work.

REFERENCES

Bekker, P., Kapteyn, A. and Wansbeek, T. Consistent Sets of Estimates for Regressions with Correlated or Uncorrelated Measurement Errors in Arbitrary Subsets of All Variables. *Econometrica*, September 1987, 1223-30.

Erickson, T. Restricting the Regression Slopes in the Errors-in-Variables Model by Bounding the Error Correlation. *Econometrica*, July 1993, 959-969.

Friedman, M. Do Old Fallacies Ever Die? *Journal of Economic Literature*, December 1992, 2129-32.

Friedman, M. and Schwartz, A. J. Alternative Approaches to Analyzing Economic Data. *American Economic Review*, March 1991, 39-49.

__________ **and** __________. *Monetary Trends in the United States and the United Kingdom: The Relation to Income, Prices, and Interest Rates, 1867-1975*. Chicago: University of Chicago Press, 1982.

Frisch, R. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo, Norway: University Institute of Economics, 1934.

Fuller, W. *Measurement Error Models*. New York: Wiley, 1987.

Haavelmo, T. The Method of Supplementary Confluent Relations Illustrated by a Study of Stock Prices. *Econometrica*, July 1938, 203-218.

Herbert, J. H. A Data Analysis and Bayesian Framework for Errors-in-Variables, in *Interfaces '88 Symposium on the Interface: Computer Science and Statistics*, edited by E. J. Wegman, D. T. Gantz and J. J. Miller. Alexandria, Virginia: American Statistical Association, 1988, 490-99.

__________. *Clean Cheap Heat: The Development of Residential Markets for Natural Gas in the United States*. New York: Praeger, 1992, 146-59.

Herbert, J. H. and Dinh, K. T. A Note on Bias from Proxy Variables with Systematic Errors. *Economics Letters*, June 1989, 207-209.

Herbert, J. H. and Kott, P. An Empirical Note on Regressions with and without a Poorly Measured Variable Using Gas Demand as a Case Study. *The Statistician*, 37, 1988, 387-91.

Kalman, R. System Identification from Noisy Data, in *Dynamical Systems II*, edited by A. R. Bednarik and L. Cesari. New York: Academic Press, 1982.

Klepper, S., Kamlet, M. S. and Frank, R. G. Regressor Diagnostics for the Errors-in-Variables Model-An Application to the Health Effects of Pollution. *Journal of Environmental Economics and Management*, May 1993, 190-211.

Klepper, S. and Leamer, E. E. Consistent Sets of Estimates for Regressions with Errors in All Variables. *Econometrica*, January 1984, 163-83.

Koopmans, T. *Linear Regression Analysis of Economic Time Series*. Haarlem, Netherlands: Netherlands Economic Institute, 1937.

Krasker, W. S. and Pratt J. W. Bounding the Effects of Proxy Variables on Regression Coefficients. *Econometrica*, May 1986, 641-55.

Kroch, E. Bounds on Specification Error Arising From Data Proxies. *Journal of Econometrics*, January 1988, 171-92.

Mendershausen, H. On the Significance of Professor Douglas' Production Function. *Econometrica*, April 1937, 143-53.

Patefield, W. M. Multivariate Linear Relationships: Maximum Likelihood Estimation and Regression Bounds. *Journal of the Royal Statistical Society*, Series B, 43, 1981, 342-52.

Reiersol, O. Confluence Analysis by Means of Instrumental Sets of Variables. *Arkiv For Matematik, Astronomi Och Fysik*, Baud 32 A, 1945, 1-119.

Ruist, E. Standard Errors of the Tilling Coefficients Used in Confluence Analysis. *Econometrica*, July 1946, 235-41.

Stone, R. *The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, 1920-1938*. Cambridge: Cambridge University Press, 1954.

Tinbergen, J. *Business Cycles in the United States of America 1919-1932*. Geneva: League of Nations, 1939.

Trost, R. P., Brown, G. D. and Woods, W. D. Measurement Error Bias in Gas Appliance Unit Energy Consumption Estimates for End Use Forecasting Models. *American Gas Association Forecasting Review*, April 1994, 24-52.