Regress to Reveal

Clopper Almon[*]

1999 January

[*] Department of Economics and Inforum, University of Maryland, College Park, MD 20742. I am indebted to Maurizio Grassini and Ralph Monaco both for discussions which led me to put this material in writing and for comments on earlier drafts. The comments of Kang Chen have added significantly to the substance.

Regress to Reveal

Suppose you have just done a regression and you *know* that the conditions for the t-values shown on your printout to be validly interpreted at Student t-statistics are *not* satisfied. In this case, what can you learn from these t-values?

I have posed this question to dozens of graduate students who had just completed the econometric training at various respectable universities. Many had some difficulty understanding so stupid a question, but once they understood it, they all replied emphatically, “Nothing.” They thereby confessed that their instruction had been so one-sided that they were missing more than half of the descriptive information about their sample which the regression gave them.. If you would have agreed with them, then you may be interested in the alternative, richer way of looking at regression presented here. This paper will show how those t-values could be converted to statistics which would present nearly the same information in an intuitively easily grasped measure, valid as a description of the sample despite the fact that the t-values could not be validly used for testing. Similar, descriptive replacements will also be offered for the standard errors of regression coefficients and F-statistics.

The emphasis on testing has led to the availability of a large battery of test statistics, most of them valid only under very special conditions. By contrast, descriptive statistics — statistics that speak to us with easily understood pictorial meaning — are an underdeveloped area. These statistics make no claim to inform us about a world beyond the sample, but they do reveal relations which exist in the sample. The notion, however, that “science” consists only of formulating hypotheses and testing them on data not consulted in their formulation has been thoroughly exploded by historians of science (Kuhn, 1962). It is just as “scientific” to explore data and look for relationships as it is to test hypotheses. Indeed, revolutionary science seems to have followed the exploratory path. In fact, it is exactly in this way that researchers generally use regression but do so with feelings of guilt. It is the purpose of this paper both to set aside the guilt and to offer several measures designed, not to test, but to *reveal* in a an intuitively comprehensible way what is happening in the data. I am sure that there are numerous further such measures, and I hope this paper stimulates further developments in this line.

Before turning to this alternative view of regression, however, it is worth reviewing the conventional view just to remind ourselves of how far it is from what we really do. In order not to bore you with a list of assumptions which you know perfectly well, let me put the review in the form of a fable. In spinning it, I had in mind more work with time-series data than with cross-section data, but it is not without relevance there also.

The Datamaker Fable

We econometricians face a body of data. Where did it come from? It was made, according our fable, by the Great Datamaker. Though we never see Datamaker, we know a lot about how he works. We know that, to make the data we are looking at, he took a matrix, $\mathbf{X}$, and a vector, β, and then generated many, many vectors, $\mathbf{y}$, by picking vectors of random numbers, $\mathbf{e}$, and calculating

$$y = X\beta + e. \tag{1}$$

He then bundled each **y** with **X** into a packet, **(X,y)**, and threw it out into the universe. One of these packets struck the Earth, burst open, and created the economy which we are studying. We have had the great good fortune to find the primordial **(X,y)**. There is no doubt about that. Our problem is to find out what β is. We know exactly what **X** is and we are perfectly sure that there was some real, true β. Though there is absolutely no chance that we will ever catch a second one of these packets, the infinitely many others are all caught elsewhere in the universe. Everyone who catches one must compute

$$b = (X'X)^{-1}X'y \tag{2}$$

and send the result to the Cosmic Information Center (CIC). The folks there -- ordinary mortals like ourselves with no direct knowledge of Datamaker's β -- will take the average of all the **b** and that average will be β. Unfortunately, confidentiality requirements preclude them from any communication back to us. So we will never know β, only our one and only **b**. Nevertheless, it is gratifying to know that we are part of their effort which will reveal to them the true β. We express our pleasure in that fact by saying that our **b** is unbiased.

Although Datamaker generally plays by the rules, he is known to sometimes play a little trick on us and include in **X** one or more variables which in fact were not used in making up **y** -- or which had a 0 coefficient in **β**. One of our particular tasks is to detect such jokes on the part of Datamaker.

Within this general fable, many details may be added. We may perhaps assume that the elements of **e** are all independent and identically distributed. That assumption allows us to compute easily the variances and covariances of all the elements of **b**. If all data catchers send along these estimates to the CIC, the average taken there will again be the true variances and covariances of **b**. We may believe further that the elements of **e** are drawn from a normal distribution. That belief allows us to deduce that the **b**'s arriving at the CIC have a multinomial normal distribution and that the ratio of an element of our **b** to our estimate of its standard error will be distributed as a Student t variable. That conclusion is very nice because it can be used to detect jokers which Datamaker may have thrown into the **X** packet. In some cases, we may believe that the elements of **e** are not independent and that we know something about the structure of the relations among them. If that knowledge is correct, it can be used to cut down on the variance of the **b**'s flying into the CIC. Though it is quite respectable to suppose that we know something about how **e** was generated, it would endanger our reputation as scientists to imagine that we know anything about β, for that would imply some economic understanding on our part.

Recently, some have supposed that Datamaker has a new trick. He makes up the elements of **y** one at a time, starting from the top, and one of the elements of the **X** matrix is just the element of **y** from the row above. Those who take this notion seriously say that they, and presumably only they, are doing "time series analysis." (Anyone working with time series data without this assumption is left homeless.) These self-styled time-series analysts devote great energy and ingenuity to determining whether or not the coefficient in **β** on this variable is equal to 1.0. We will not pursue this school further save to note that its results seem to depend very heavily on knowing exactly how Datamaker works.

Seeds of Doubt

I have the greatest possible admiration of the ingenuity and beauty of the mathematical derivations based on the Datamaker fable. The derivations of the distributions of the regression coefficients, of t- and F-statistics are marvelous. Von Neumann's derivation of the distribution of his δ (on which the Durbin-Watson statistic is based) is, for me, miraculous. I am awed by the thoroughness of theoretical econometricians in working out the consequences of various assumptions. For years, the sheer beauty of the derivations blinded me to the basic fact that the Datamaker fable has little connection with what I am doing as an applied econometrician working with time series data. That is not to say that there may not be cases where the Datamaker fable may be entirely appropriate, as in the analysis of repeatable, controlled experiments, where one knows exactly what has changed from one experiment to another.

But that is not what I am doing as a builder of econometric models. I have one set of data on the American economy in the 1990's and there is no chance that I will ever get a second set with only certain known policy changes. Furthermore, the process generating the data is vastly more complex than any equation I can write down, though I may have some insight into it, and I may try to capture that insight in the equation. I do not, however, believe for one second that I know the full $\mathbf{X}$ matrix nor, indeed, that there is any true β. As a builder of economic models, I am just looking for rough but workable approximations of a vastly complicated reality. I model consumption of ice cream with income and relative prices. But you buy ice cream; you know that how much you consume depends on how hot the weather is, on whether or not you or your children have milk allergies or philosophical positions about animal-derived food, on what kind of diet you may be on, and on how loudly the children are howling in the back seat. Price? Income? Hardly. The one thing I am relatively sure of is that *there is no true equation of the form I am fitting*. All claims about the **b**'s that I calculate being "unbiased" or "consistent" "estimates" of your true parameters seem pretty meaningless. Most econometric theory seems, in the end, to be about how to make unbiased, consistent, efficient estimates of non-existent parameters.

I am surely not alone in doubting the appropriateness of the fable to what we are doing. Poirier (1988, p. 132) notes "Such parameters need not 'exist' in the external world, but only in the minds of researchers." He finds them regarded as anything from metaphysical mental constructs to "waste products" of prediction. Leamer (1983), Sims (1996) and others have expressed various doubts on this point. McCloskey and Ziliak (1996), in their criticism of the profession for all too often forgetting the difference between "statistical significance" and "economic significance," rightly observe "Essentially no one believes a finding of statistical significance or insignificance." Keuzenkamp and Magnus (1995) have offered a handsome reward to anyone who can produce one point on which the opinion of the profession has been changed by a significance test. If significance tests have, rightly, lost all persuasive power, perhaps we should look for other, less presumptuous ways of presenting the information about the sample that the test statistics do, actually, contain.

These arguments certainly do not mean that I have no use for regression. Quite the contrary. I find it an indispensable tool in economic modeling, which, despite the criticism of recent years, remains the only way that I know to test my understanding of the economy and to put together pieces of understanding into a coherent whole. In modeling, I am looking for a workable summary of extraordinarily complicated economic behavior. I find it helpful to admit that complexity, not to gloss over it by the Datamaker assumption. I do not regard, however, the regression coefficients as *estimates* of anything. They are just a sort of summary statistic of the data. My concern is not to reject regression but to make it speak in terms that are easily understood without invoking Datamaker.

I will use the word “metaphysical” to describe a statement which relies on the Datamaker fable for its meaning. In doing so, I intend no offense to the science of metaphysics, nor indeed, to say that the statement is vague or unreal. In reading Aristotle’s *Metaphysics,* however, it struck me that his unmoved mover and Datamaker might be of similar substance. I only wish to say that these statements rely for their validity on the existence of a reality beyond anything we can observe — that they go “beyond nature.” They may assume, fore example, the existence of a true **β**, a transcendent reality beyond our powers of observation. I will use the word “factual” to describe a statement that does not invoke the unobservable; its meaning is intuitively clear without the fable. If I say that I have regressed **y** on **X** with ordinary least squares and the result was **b**, that is a “factual” statement. If I say that **b** is an unbiased estimate of **β**, that is a “metaphysical” one. If I say that the standard deviation of the residuals is 16, that is a factual statement. If I add that the 16 is an unbiased estimate of the standard deviation of the normal distribution from which the elements of **e** were drawn, that is a metaphysical statement.

"That was easy," you may say, "but what about the standard errors of the regression coefficients, the t-statistics and the F-statistics. Aren't they all inextricably bound up with the fable?" Indeed, the names we give these measures are justified only by the fable. Without the fable, these particular measures are virtually incomprehensible; they make no intuitive, pictorial sense. In this sense, I will call them also “metaphysical” statistics. In other words, if a particular measure can be used to make a meaningful factual statement, I call it a “factual statistic”; if is well adapted only for making metaphysical statements, I will call it a “metaphysical statistic.”

How to replace the metaphysical statistics to which we are all accustomed with factual statistics which convey essentially the same information but in an form meaningful without Datamaker is the subject of the rest of this note.

Factual Loss Limits and Metaphysical Standard Errors

Let us begin with "standard error of the regression coefficient." There must be conceivably more than one of something for the concept of standard error to make sense. The very idea that the regression coefficients have standard errors depends upon there being, at least potentially, many **b** vectors. When we are working with economic time series and trying to estimate equations for, say, the U.S. economy in the 1970 - 1998 period, only the Datamaker assumption that many **(X,y)** packets are cast off into the universe can supply the multiplicity of **b**'s, for we shall certainly not see these years re-run with just the "errors" changed. We are essentially working with the whole population; and, without, Datamaker, there is no meaningful standard deviation of the regression coefficients. If you ask me, "What was the average value of the Treasury bill rate in the 1980's?" you expect an answer like, "8.8 percent." If I add, "and the standard deviation of that mean is .44 on the assumption that our 1980's were a random sample from all possible 1980's," you are likely to mutter, "No, no, I just wanted to know about the 1980's as they really were," and think me some kind of lunatic. The mean that I gave you, however, was just the regression coefficient of the Treasury bill rate on a series of 1's, and I thought — in my lunatic way — that you would want to know its standard deviation. Standard errors of regression coefficients on economic time series data are all more or less in the same class with my lunatic answer about the standard deviation of the mean. The regression coefficients themselves, however, are useful descriptive statistics.

With random samples of cross-section data, matters are a bit more favorable to the usual interpretation, for we can conceivably draw multiple random samples and compute **b** and the 95-percent confidence interval for each and reasonably expect that about 95 percent of these intervals will include the **b** that would be found by running *exactly the same regression* on the whole population. Even in this case, however, the confidence intervals tell us nothing about what to expect if another variable is added to the regression. Only the belief that we know the full **X** matrix used by Datamaker enables us to make any statement that transcends the particular choice of variables we have made. And should we find that one of the variables is "insignificant" and rerun the regression without it, the standard errors the program gives may be utterly misleading, for we may have made a type II error in throwing out the variable.

We may, however, in any case ask What is the factual content of the number that is usually reported as the standard error of a regression coefficient? That statistic is really just giving us information about how rapidly the sum of squared residuals (SSR) rises as that regression coefficient is moved away from its least-squares value and the other regressions coefficients change to compensate, as best they can, for that movement. To be more precise, let us divide the **X** matrix vertically into two parts, $\mathbf{X_1}$ and $\mathbf{X_2}$, where $\mathbf{X_2}$ contains only a single variable and $\mathbf{X_1}$ contains all the others. Similarly, we divide up the **b** vector between $\mathbf{b_1}$ and $\mathbf{b_2}$ and then define the vector **r** of residuals by

$$\boldsymbol{r = y - (X_1 b_1 + X_2 b_2).} \qquad \textbf{(3)}$$

Now let us take hold directly of $\mathbf{b_2}$ and move it about, but always changing $\mathbf{b_1}$ so as to minimize the sum of squared residuals. Thus the SSR becomes a function of the $\mathbf{b_2}$ we choose, which we may call SSR(b_2). We can easily write it down:

$$\boldsymbol{SSR(b_2) = (y - X_1(X_1'X_1)^{-1}X_1'(y - X_2b_2))'(y - X_1(X_1'X_1)^{-1}X_1'(y -} \qquad \textbf{(4)}$$

Expanding and simplifying gives:

$$\boldsymbol{SSR(b_2) = \left(y'y - y'X_1(X_1'X_1)^{-1}X_1'y\right) + 2\Big(X_2'X_1(X_1'X_1)^{-1}X_1'y - X_2'y} \\ \boldsymbol{+ \Big(X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\Big)b_2^2.} \qquad \textbf{(5)}$$

To see what familiar friends those long matrix products really are, let us write out the normal equations for the regression of **y** on **X** and the simultaneous inversion of **X'X** to create its inverse, **S**. They are just

$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} b_1 & S_{11} & S_{12} \\ b_2 & S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} X_1'y & I_{11} & 0 \\ X_2'y & 0 & 1 \end{pmatrix} \tag{6}$$

If we now proceed with the Gauss-Jordan elimination process to the point just before the final pivot operation to determine $\mathbf{b_2}$, we have

$$\begin{gathered} \begin{pmatrix} I & (X_1'X_1)^{-1}X_1'X_2 \\ 0 & X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2 \end{pmatrix} \begin{pmatrix} b_1 & S_{11} & S_{12} \\ b_2 & S_{21} & S_{22} \end{pmatrix} = \\ \begin{pmatrix} (X_1'X_1)^{-1}X_1'y & (X_1'X_1)^{-1} & 0 \\ X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y & -X_2'X_1(X_1'X_1)^{-1} & 1 \end{pmatrix} \end{gathered} \tag{7}$$

Let us denote the matrix on the left here by **A** so that a_{22} is the element in the lower right corner. By looking back at equation (5), we see that this element is precisely the coefficient on b_2^2 in (5). Moreover, the coefficient of b_2 in equation (5) is just -2 times the expression in the lower left corner of the matrix on the right of equation (7). Furthermore, the last step of the Gauss-Jordan process will divide this element by a_{22} to produce $\mathbf{b_2}^*$, the least-squares value of $\mathbf{b_2}$. The first term on the right of (5) is just SSR_1, the SSR resulting from the regression of **y** on just $\mathbf{X_1}$. Thus, equation (5) can be written as

$$SSR(b_2) = SSR_1 - 2a_{22}b_2^*b_2 + a_{22}b_2^2. \tag{8}$$

Now the final pivot operation for solving equation (6), the step following that shown in equation (7), will involve dividing the 1 in the lower right corner of the matrix on the right in (7) by a_{22} to get s_{22}, the diagonal element of the $\mathbf{(X'X)^{-1}}$ matrix corresponding to the coefficient we are moving. Thus, equation (8) can be written as

$$SSR(b_2) = SSR_1 - 2b_2^*b_2/s_{22} + b_2^2/s_{22}. \tag{9}$$

Setting $\mathbf{b_2 = b_2^*}$, we find for the SSR for the full least squares regression, which we may call SSR^*,

$$SSR^* = SSR_1 - b_2^{*2}/s_{22}. \tag{10}$$

If we now introduce δ as the deviation of $\mathbf{b_2}$ from its least-squares value and substitute $\mathbf{b_2 = b_2^* + \delta}$ into (9), it becomes, after simplification,

$$SSR(\delta) = SSR^* + \delta^2/s_{22}. \tag{11}$$

Suppose now that we ask, How far from its least-squares value can we move $\mathbf{b_2}$ before the SSR for the whole equation would increase by more than λ percent? The answer, quickly deduced from equation (11), is

$$\delta = \sqrt{.01\lambda \cdot SSR^* \cdot s_{22}} \tag{12}$$

If for example, we picked λ = 5, then we would have for what we might call the "five-percent loss limit" on $\mathbf{b_2}$

$$\boldsymbol{\delta} = \sqrt{\mathbf{.05}\cdot \boldsymbol{SSR}^{*}\cdot \boldsymbol{s}_{22}}. \qquad \mathbf{(13)}$$

Now if a regression has 20 degrees of freedom, what is the "standard error" of b_2 by the usual calculations? Exactly the δ given by equation (13). For 20 degrees of freedom, the metaphysical "standard error" of the regression coefficient as printed out by the computer is, factually speaking, its five-percent loss limit. If there were 100 degrees of freedom, the metaphysical "standard error" would be the factual 1 percent loss limit, and so on.

For the calculated number to really be a standard error, a whole host of assumptions must be valid. Firstly and most unlikely, there must be a true equation of exactly the form we are estimating. Secondly, we must be sure that we know *a priori* what X is. If we have done any previous regression and discarded some variables on the ground that their t-statistics were insignificant, then through this pre-test we have admitted that we do not know what the true X is; and our present estimates of **β** are biased (because we may have made a Type II error and thrown out a variable which belongs in the equation), and the standard errors are more or less meaningless. (This point is eloquently made by Fomby *et al.* [1984, p. 130].) Thirdly, **X** must be non-stochastic. Fourthly, the errors must be uncorrelated with one another. Fifthly, they must all have the same variance. Sixthly, if the standard error is to be used to calculate a valid t-statistic, the errors must also be normal. By contrast, the factual loss-limit statement is always valid. If you change a regression coefficient by its five-percent loss limit and recompute the others by least squares, the SSR will for sure and certain go up by five percent. That is just a fact. (Some of these conditions can be relaxed a bit for large samples, but that fact hardly helps the worker who has twenty years of historical data with which to fit his equation. He can't go back further because the structure of the economy, the β, would have almost certainly changed and he can't go into the future, because those data don't yet exist.)

The second of these conditions almost eliminates the valid use of classical statistical methods in economics. These methods are aimed at estimating parameters or testing hypotheses when the correct specification of the equation is known. But the notion that economic theory will tell us what variables to put into an equation and the form of the equation is almost always simply laughable. If we are to find equations with acceptable fits, we have to rely on our own explorations of the data or on the empirical work of others. That reliance totally invalidates the classical statistical tests and sampling properties. Loss limits, being purely factual statements, remain perfectly valid no matter how much we have explored the data. They are, of course, descriptive only of the sample and do not make any claim on a wider applicability.

What happens to loss limits and standard errors as the sample size increases? Suppose for example that we were able to double the sample and that it just so happened that the additional observations turned out to look, one for one, exactly like the first set. The loss limits will be unaffected by such a doubling of the sample. The standard errors will all shrink by a factor of 1/√2. Large sample studies nearly always have tiny standard errors and huge t-statistics. Isn't that nice? Their loss limits, however, are not necessarily very different from those of regressions on a much smaller sample. Isn't that a drawback for the use of loss limits? Is there any way in which a factual statement can express the superiority of the large sample? In my view, the factual statement is simply the sample size and its structure. I am very leery of tiny standard errors in large-sample studies, because the large sample is just as sensitive as the small to errors in specifying the X matrix. The fact that some regression coefficient is ten times its standard error is supposed to make me very confident of its sign. But the truth is that I don't really know what the **X** matrix should include. After I have done my best you may come along and suggest a new variable. When I throw it into the regression, lo, the sign changes on the variable whose t-statistic was 10. My confidence in my metaphysical knowledge shattered, I decide to stick to factual statements next time.

Since the loss limit statement is so much more factual than the standard error statement, one might well ask that a regression program display loss limits. The G regression program available on the Internet at inforumweb.umd.edu does so. If you give the command "ll 5", then after the next regression you will see the 5 percent loss limits for each coefficient.

Factual Mexvals and Metaphysical t-Statistics

Most regression programs report the t-statistics for each regression coefficients. Their main use is in deciding whether or not the variable is one of the jokers that Datamaker slipped into the packet. Their validity is subject to all the conditions we have just enumerated for the standard errors. If we have the slightest doubt about their validity we can ask the factual statement, How much does the SSR increase if we drop this variable? The answer is immediately clear from equation (10). It goes up by b_2^{*2}/s_{22}. A convenient way to express the answer is to ask by what percent the standard error of estimate goes up when the variable is eliminated and all others adjust to compensate as best they can for the elimination. We may call this measure the marginal explanatory value, or mexval, of the variable. If we denote it by m in general and by m_2 for the particular case we have been developing, then

$$m_2 = 100\left(\sqrt{\frac{SSR^* + b_2^{*2}/s_{22}}{SSR^*}} - 1\right) \qquad \textbf{(14)}$$

The t-statistic is

$$t_2 = \frac{b_2^*}{\sqrt{s_{22}\cdot SSR^*/(T-n)}} \qquad \textbf{(15)}$$

so if your software fails to compute the mexvals, you can do so yourself by the equation

$$m = 100\left(\sqrt{1 + \frac{t^2}{(T-n)}} - 1\right) \qquad \textbf{(16)}$$

where T is the number of observations and n is the number of parameters estimated. (You might also consider switching to the G software or demanding that the makers of your software put in mexvals.)

Just as the relation between the loss limits and the standard errors depended on the degrees of freedom in the equation, so does the relation between mexvals and t values. Which one is telling you what you want to know? Consider an equation with a variable that has a t-statistic of 3. If that equation has 10 degrees of freedom, eliminating the variable will wreak havoc with the fit: mexval = 40. If the equation has 1000 degrees of freedom, though the variable is somewhat more "significant" by the t-test, eliminating it will have little effect on the fit: mexval = .45. As a non-believer in Datamaker, I find the mexvals to be telling me exactly what I want to know in the two cases but the t-statistics to be tricky to compare.

Factual Derivatives and Metaphysical Covariances

What sort of factual statements correspond to the covariances of regression coefficients? If we return to equation (3) and ask how **b**$_1$ changes to compensate for changes in **b**$_2$, we find

$$\boldsymbol{b}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\boldsymbol{b}_2) = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\boldsymbol{b}_2. \tag{17}$$

The matrix (actually, it is a vector) which is multiplied by b_2 in the last term of the right side of this equation is the derivative of **b**$_1$ with respect to b_2. Now note in equation (7) that if we carry the Gauss-Jordan pivoting process to its conclusion we will have

$$\boldsymbol{S}_{12} = -s_{22}(X_1'X_1)^{-1}X_1'X_2. \tag{18}$$

Note the similarity to the coefficient of **b**$_2$ on the extreme right of (17). Recalling that the variance-covariance matrix by the usual formula is s^2S, we see that if we divide each of its columns by the diagonal element in that column, we obtain a matrix whose j^{th} column shows the derivatives of all the regression coefficients as b_j is independently varied and all the others are varied to maintain as good a fit as possible with the given b_j. This matrix of derivatives is the factual way of interpreting the information contained in the metaphysical variance-covariance matrix. In factual terms, the variance-covariance matrix is showing us how sensitive the other regression coefficients are to the value chosen for any one. One could, of course, also multiply each column of this derivative matrix by, say, the five-percent loss limit for the corresponding variable to see how far each of the other regression coefficients would move if a given one were moved out to its five-percent loss limit.

Factual Normalized Residuals and Metaphysical F Statistics

If a regression is computed by successive Gaussian pivots, it is little extra work to carry one more row which will give in the diagonal element the SSR after each pivot. If these numbers are saved, they can be used for printing at the end of the regression the F statistics for testing, under the usual Datamaker assumptions, the significance of the last variable, the last two, the last three, and so on through the whole equation. (If your software gives only one F, the one for the whole equation, change to G or demand an improvement.) These F's are, of course, designed for making metaphysical statements about significance. The same information can be conveyed factually by simply showing the SSR for each stage, or by expressing each of them as a ratio to the SSR when all variables have been included. In the G program, these ratios are called "Normalized residuals" because they have been normalized by the last one. They are routinely shown by G and are helpful for judging the usefulness of a group of variables, especially if the group is placed at the end of the list of regressors. These ratios are, of course, simply factual statements without metaphysical overtones.

Other Factual Statistics

A number of other standard statistics are factual in nature. For example, the ρ or autocorrelation coefficient of the residuals has a simple intuitive meaning as the regression coefficient of the residual on its lagged value, the tendency of the equation to go on making the same mistake. Putting the same information in the form of a Durbin-Watson statistic takes away the intuitive interpretation and raises the suspicion that one has in mind making some metaphysical statement about how Datamaker drew the **e** vector. The mean absolute percentage error is a factual statistic, as are the elasticities of the various variables evaluated at the means of the observed values. The leverage vector, used in detecting outlying observations, is simply the derivative of the predicted value of each observation with respect to its observed value. It is also factual. Beta coefficients, which express the regression coefficients in units of standard deviations of the dependent and independent variables, are likewise factual.

Data Mining, Factual Statistics, Judging Regressions and Prior Information

Exploring the data with regression analysis certainly invalidates the metaphysical test statistics. It is therefore often held to be reprehensible and is referred to in pejorative tones as "data mining" or "data snooping." Let me say plainly that I think that it is the responsibility of the researcher to explore the data thoroughly. Isn't that what makes one an expert on a subject? Isn't that precisely what the researcher is getting paid for? Have you ever, on looking at someone else's regression, asked, "Did you try so and so?" If so, you explicitly recommended data exploration. Indeed, if we are not allowed to learn about the real world by looking at data, how then are we supposed to learn about it? From other researchers who have also not looked at their data?

So if the researcher has done a thorough job, the data is completely mined and the conventional test statistics utterly misleading. *The factual statistics, however, remain perfectly valid for the sample.*

Does this attitude open the spillways to all manner of junk regressions? Not at all. The next step after estimating an equation is to use it in a model. To do so implies that we expect that the relations found by the equation will continue to hold in the future or at least would have held in the past even if some of the independent variables had been different. That expectation gives us a number of ways to judge an equation. In Almon [1994], there is a checklist of such criteria which have nothing to do with test statistics. They include accounting for important influences, parsimony, appropriate dimensions, reasonable attention to cointegration, adequate allowance for lags, plausible parameter values, stability of coefficients when the sample period is changed, satisfactory fit, and several others not easily explained out of context. The leverage variable should be examined to detect outlying observations and those observations considered carefully. Indeed, the notion that all an equation needs is a high R^2 and significant t-statistics will certainly admit more junk equations than do these criteria.

Since plausibility of regression coefficients is a primary concern for me, one might suppose that I would use (or at least advocate that others use) Bayesian regression. But the Bayesian position, just as much as the classical position, involves assuming that there are true parameters. One who holds that there are no true parameters needs a procedure closer to the emphasis on regression coefficients as summaries of data. If we want the parameters of an equation to satisfy approximately some linear constraint -- the simplest being that the parameter should have a certain value -- but the regression refuses to give "nice" values, we can just make up *artificial data* which would be fit perfectly by any equation whose parameters satisfy the constraint. We then combine this artificial data with the natural data in proportions to give a balance between our desires that the equation fit both the natural and the artificial data. A good regression package can make it extremely easy to use these "soft" or "stochastic" constraints without any appreciable increase in the time required for the regression computations. As with Bayesian regression, use of this procedure obligates us, of course, to report the use of the artificial as well as the natural data. The use of the artificial data affects the loss limits, mexvals, and normalized residuals, for in their calculation the artificial data is just as much data as is the natural data.

Conclusion

It has proven possible to give factual alternatives to all the common metaphysical statistics. In reporting results from regression analysis, you do not have to make metaphysical statements that you don't believe. You can convey the same information to your readers with purely factual statistics. These statistics can easily be incorporated into regression programs, as they already are in the G program. By a de-emphasis of testing and an increased emphasis on economic measurement and interpretation, I hope that they will contribute to putting both the *econ* and the *metrics* back into econometrics.

References

Almon, Clopper, *The Craft of Economic Modeling*, 3rd ed. Part I. College Park, MD, (P.O. Box 451) Interindustry Economic Research Fund, 1994

Fomby, Thomas B., R. Carter Hill, Stanley R. Johnson, *Advanced Econometric Methods,* New York, Berlin, Heidelberg, Tokyo: Springer Verlag, 1984

Kuhn, Thomas S. *The Structure of Scientific Revolutions,* Chicago, IL: University of Chicago Press, 1st Ed. 1962, 3rd Ed. 1996)

Keuzenkamp, Hugo A., Jan R. Magnus, “On tests and significance in econometrics”, *Journal of Econometrics*, 67 (1995) 5-24.

Leamer, Edward E., "Let's Take the Con out of Econometrics," *American Economic Review*, March 1983, *73:*1, 31 - 43

McCloskey, Deirdre N. and Ziliak, Stephen T., “The Standard Error of Regressions,” *Journal of Economic Literature*, March 1996, *34:*1, 97-114

Poirier, Dale J., Frequentist and Subjectivist Perspectives on the Problems of Model Building in Economics, *Journal of Economic Perspectives*, 1988, *2:*1, 121-144.

Sims, Christopher A., Macroeconomics and Methodology, *Journal of Economic Perspectives*, 1996, *10:*2, 105-120.