VOLUME 15 NUMBER 3/4 MARCH/APRIL 1990 ISSN 0360-5442

TECHNOLOGIES · RESOURCES · RESERVES · DEMAND

IMPACT · CONSERVATION · MANAGEMENT · POLICY

The International Journal

SPECIAL ISSUE

ENGINEERING–ECONOMIC MODELING: ENERGY SYSTEMS

Part I

Guest Editors:

John P. Weyant and Thomas A. Kuczmowski

PERGAMON PRESS

Oxford · New York · Beijing · Frankfurt
São Paulo · Sydney · Tokyo · Toronto

Energy Vol. 15, No. 3/4, pp. 213–236, 1990
Printed in Great Britain

0360-5442/90 $3.00 + 0.00
Pergamon Press plc

2.1. ECONOMETRIC TECHNIQUES: THEORY VERSUS PRACTICE

PAUL J. WERBOS

Energy Information Administration+
Department of Energy, Washington D.C.
U.S.A.

Abstract - This paper introduces the basic concepts used in econometric modeling, and describes five prescriptions to avoid common real-world pitfalls in that style of modeling. The paper begins by comparing econometric modeling with other forms of modeling used in energy modeling and engineering. It describes what an econometric model is, and how to build one. It then gives a detailed explanation of many facets of the five prescriptions: pay attention to uncertainty; don't expect a free lunch when devising specifications; pay attention to prior information; don't expect to draw conclusions without adequate data; and check the historical track record of your model. The issues of generalization and robustness over time receive special attention; they are important in practice, and subtle in theory. Finally, the paper discusses model development in practice, building upon experience with PURHAPS, a model I developed for the Energy Information Administration (EIA).

1. BACKGROUND

Economic theory, in the United States, usually begins with simplifying assumptions like free markets, perfect competition, no externalities, and perfect foresight. After years of study, the advanced student is told how to modify this theory to address real problems in the real world, which are often quite different from the theory in important ways. Some students never quite make the adjustment.

Econometrics is very similar. This paper will introduce the novice to the basic assumptions and methods of econometrics, and then discuss problems which come up in modifying the theory to fit the real world.

Broadly speaking, there is no sharp dividing line between econometric models, engineering process models, statistical models, simple time-series models, systems dynamics models, etc. All these types of models are systems of equations designed to forecast or simulate whatever we want to forecast or simulate. The real difference lies in how we obtain information or parameters to plug into the models.

Some classes of models tend to rely on a priori information or indirect information about what we are forecasting; models of this sort include "pure" process models, classical systems-

---

+This paper expresses the views of the author, not those of EIA or DOE, though it was reviewed at EIA prior to submission. As this paper goes to press, the author's address has changed to: Room 1151, NSF, Washington D.C. 20550.

dynamic models and expert systems. Other classes are based on empirical data about exactly the kind of variables we are trying to forecast; this includes econometric models, time-series models, statistical models, (identified) control-theory models, and artificial neural networks.

In the energy business, the a priori models tends to be very complex, because people include lots of lower-level detail to create a feeling of earthy realism; however, the parameters are usually based on judgment or guessing, and it is hard to be sure the model will track actual trends. The good empirical models tend to be simpler, because they are usually limited to variables which are observed on a time-series basis; however, they are strongly rooted in empirical reality, if done right.

The a priori models sometimes seem easier to understand, at first, because they mimic concrete, well-known engineering processes (at least in part); however, because they contain so much detail, it is not always easy to know what causes the bottom-line forecasts to come out as they do, and the role of human behavior is often neglected or oversimplified. Empirically-based models are the reverse: the overall behavior is easier to understand, but the detailed reasons behind the trends -- both historically and in the forecast -- may require further analysis. A good researcher will learn how to combine both prior information and empirical information into a model, as this paper will discuss.

The relations between different kinds of empirical model are subtler.

On some level, there is no real difference between a statistical model, an econometric model, and a model developed by using the identification techniques of control theory; all three rely on the same core of theory, which this paper will discuss. "Simple time-series models" and artificial neural networks depend on the same theory as well, but they try to automate the process of coming up with a functional form; in effect, they assume that the user does not really understand the structure of the system he is studying, so that a computer can do the job as well as a person. This is a good assumption in some cases (as in recognizing patterns among thousands of variables which no one fully assimilates into his or her intuition), and a poor assumption in others (as in the study of physical phenomena for which the dynamics are well-understood).

2. GOALS OF THIS PAPER

In the United States and many other nations, econometrics is a major academic discipline, based on the idea that a careful analysis of historical data can be a good starting point for analyzing or projecting the future. Like any major discipline, econometrics has a long history, full of false starts, new perspectives, and hundreds of applications, some good and some bad.

This paper will present those concepts and rules of thumb which we have found most important, in practice, in a government organization concerned about the quality of its forecasts. No one can expect to become a first-class econometrician after reading one article; there are simply too many tricks and traps to learn. However, we will try to explain the key concepts, and cite books which elaborate on their application. We also hope to pinpoint those misunderstandings which are common among experienced practitioners, and we apologize to them that there is not enough space here to explain all the details. Unfortunately, these misunderstandings have often led to the creation of models which totally misrepresent the dynamics of the variables which they are supposed to predict.

This chapter will begin by saying what an econometric model is and how -- mechanically -- to build one. Next, it will discuss five major prescriptions for the correct use of econometric tools. Then it will discuss the use of these tools in practice at the Energy Information Administration (EIA). It will conclude with a very quick overview of the PURHAPS model, one of the econometric models I have developed for EIA.

3. WHAT IS AN ECONOMETRIC MODEL?

Strictly speaking, an econometric model is no different from any other forecasting model -- it is made up of any set of equations or formulas which can be used to predict the future. For example, consider the following simple model to predict population:

$$POP(t+1) = c \cdot POP(t) \quad (1)$$

This says that population in year t+1 is equal to a constant "c" multiplied by the population in year t. If you obtain your estimate of the constant c by asking your boss what c should be, or by studying textbooks on theology or ideology, then we would call this a judgmental model. If you obtain your estimate of c from small-scale case studies of controlled populations, we might call this an engineering model. If you have an historical time-series of data on population, for the state or nation whose population you are forecasting, and if you estimate c from that time-series in a rigorous way, then we would call this an econometric model. In principle, then, there is no such thing as an econometric model; there are only econometric methods for estimating parameters such as "c" in general models. A pure econometric model is simply a general model, in which all of the parameters have been estimated by econometric methods, based on empirical data.

Econometric methods were initially developed for use in economic forecasting. However, there is nothing in our discussion which will restrict their use to economics. Econometric methods have often been applied directly to forecasting social and political systems (Werbos, 1974; Werbos, 1977; Werbos and Titus, 1978). Human minds and computers which truly imitate human minds must also have a built-in capability to learn cause-and-effect relations by somehow analyzing a time-series of sensory experience; we have shown how econometric methods may be embodied directly into the wiring of such systems (Werbos, 1987a; Werbos, 1986a).

In general, people who use historical data or trends or track records to help them make decisions are making inferences about cause and effect. Like it or not, they are engaged in a form of statistical inference. Even if they say they are merely testing an hypothesis, or a relation, and not formulating a model, the fact is that they are estimating a model; the potential for error and uncertainty is merely less visible and harder to correct when they deny this fact. (Of course, some managers would prefer to hide such uncertainties from their superiors. If a superior really cannot understand econometrics, there is an art to using econometrics properly and then translating the results back into English, using graphics and discussions of percentage growth rates and historical analogues.)

4. HOW CAN AN ECONOMETRIC MODEL BE BUILT?

The first stage in building a model is to review the available data and concepts, as we will discuss further in the section on "Practice".

Next one must choose a computer package to work in, to implement the econometric methods. EIA generally prefers to use the SAS package (SAS, 1985a; SAS, 1985b) on its large computer, because of its superior flexibility and data-handling capabilities; however, Troll (1981) has also been used, because of the sophisticated econometric tools it contains (some developed under contract to us). On microcomputers, SAS is also available, but is relatively expensive at present; Lotus is widely used, and new packages from Wharton Econometric Forecasting Associates (of Philadelphia) and elsewhere may be used more in the future. We use SAS to estimate parameters such as "c" in equation 1, and to evaluate the overall degree of fit of equations such as 1 and alternatives to 1; then, when all the equations are estimated and selected, we usually program the forecasting itself in FORTRAN. Actually, SAS and Troll have

the capability to simulate the model -- to generate forecasts - as well; we have used that capability only rarely, because our models have usually been too big to fit into those systems.

The next step is simply to use the package chosen. To estimate equation 1, for example, you would first locate a time-series of data on the variable POP, and load it into a SAS dataset using the SAS command DATA (SAS, 1985a). Then you would use a SAS command such as GLM (SAS, 1985b) to estimate "c" in equation 1, and to evaluate the error which this equation would have led to in forecasting the past. You could also use SAS to estimate alternative equations, and their errors, and you could select between equations based on their error. At that point, you have the equation, and you need only code it into a forecasting program.

The most common way to estimate a complex model, in econometrics, is to use "regression" or "least squares." Using regression, we estimate each equation of the model separately, one after another. For each equation, the regression command finds those values for the parameters which lead to the smallest possible error over the historical period you have data for. "Error" is defined as the sum over all observations of the square of (actual minus predicted). Regression also reports what the error is for the equation as estimated.

In actuality, most computer packages have two main regression commands available -- a linear regression command and a nonlinear regression command. (See Wonnacott and Wonnacott, 1977, Chapters 13 and 15, for more explanation.) To avoid complications, most economists use linear models such as the following two-equation model:

$$Y(t) = a \cdot Y(t\text{-}1) + b1 \cdot X_1(t) + \ldots + bn \cdot X_n(t) + c \qquad (2a)$$

$$Z(t) = c1 \cdot Y(t) + c2 \cdot Y(t\text{-}1) \qquad (2b)$$

In equation 2a, Y(t) is the "dependent variable" -- the variable being predicted in that equation. Y(t-1) and $X_1(t)$ through $X_n(t)$ are the independent variables of that equation.

The term "c" is the "constant term" or "intercept;" note that equation 2b has no intercept. The parameters of the model are the constants a, b1 through bn, c, c1 and c2. The "endogenous variables" -- Y and Z -- are the variables being predicted somewhere in the model. Y(t-1) is a "lagged endogenous variable" (because Y is endogenous and because t-1 represents a "lagged" value, a previous year's value.) X_1 through X_n are "exogenous" because they are not endogenous.

This example is linear, because the dependent variable in every equation is predicted as a linear combination ("weighted sum") of the independent variables, plus an optional constant term. To estimate each equation in SAS, you need only use the linear regression command (GLM or something similar) once for each equation. You can be sure of quick results, and you do not have to give an initial guess for the values of the parameters. Each time, you only have to tell SAS the name of the dependent variable and the names of the independent variables. You also have to tell SAS whether you want a constant term in the equation, and whether you want SAS to print out all the diagnostic statistics anyone has ever thought of.

At first glance, equation 2a may appear somewhat abstract and unrealistic. Economic relations in the real world are often more complex. For example, even in a simple model of fuel oil demand (QOIL) as a function of residual oil prices one would not want to use the simple equation:

$$QOIL = a \cdot PRESID + b \cdot PDIST + c \cdot DISTSHARE \qquad (3)$$

If you used this equation, by regressing QOIL on PRESID, PDIST, and DISTSHARE, you would expect to find that "a" and "b" are estimated as negative numbers, expressing the idea that higher prices lead to lower demand. However, with "a" and "b" negative, there will always exist a price so large that demand becomes negative, which is an absurd forecast. Likewise, the effect of

changes in PDIST should depend on how large the distillate share is; PDIST and DISTSHARE have an "interaction effect." For these reasons, a better specification would be:

$$LOG(QOIL) = a + b \cdot LOG(POIL) \tag{4a}$$

$$POIL = PDIST \cdot DISTSHARE + (1 - DISTSHARE) \cdot PRESID \tag{4b}$$

Equation 4a says that QOIL is a function of the weighted average price of fuel oil, POIL. It says that a given percentage change on POIL leads to a proportionate percentage change in QOIL; the factor of proportionality is just "b", the price elasticity of demand. (To see this, differentiate 4a or see Wonnacott and Wonnacott, 1977, Section 15-3). Equation 4a can be estimated easily in SAS by first using a DATA step to calculate:

$$LOGQOIL = LOG(QOIL)$$

$$LOGPOIL = LOG(PDIST \cdot DISTSHARE + (1 - DISTSHARE) \cdot PRESID,$$

and then calling the regression command and asking it to regress LOGQOIL on LOGPOIL. Equation 4b contains no parameters at all to estimate; it is called an "accounting identity" (as opposed to the "behavioral equation" 4a). Equation 4a is linear in the parameters a and b, but not in the original variables QOIL and POIL. Most econometric models are linear in the parameters but not in the original variables. Most of them also use tricks like the above to express economic relationships.

If equation 4a had actually been nonlinear in its parameters, then nonlinear regression could have been used. Nonlinear regression requires a lot more care and patience, depending on what computer package you use, but there is usually a way to make it work. Likewise, there are alternatives to regression which would require you to estimate the entire model as a system, together; to use these alternatives, you would have to type both equations into a single model file or command block.

Aside form their linearity, the models in equations 2 and equations 4 have two other simplifying features. First, they are "recursive". In economics, this means that they are really just simple formulas; you can calculate a forecast by plugging in values for the exogenous variables and lagged endogenous variables, and using the equations one after another like a formula or a recipe. Most econometric models are actually simultaneous, as in the following example:

$$LOG(SUPPLY) = a + b \cdot LOG(GNP) + c \cdot LOG(PRICE) \tag{5a}$$

$$LOG(DEMAND) = d + e \cdot LOG(GNP) + f \cdot LOG(PRICE) \tag{5b}$$

$$SUPPLY = DEMAND,$$

where GNP is exogenous and where the model is used to forecast a PRICE that makes SUPPLY and DEMAND balance. to make a forecast, you cannot jus plug in GNP and PRICE on the right-hand side; you cannot, because you don't yet know that PRICE is. You have to solve this system of equations, as a set of three simultaneous equations in three unknowns. In fact, if you insert these three equations into Troll (1981), Troll will take care of this problem and give you a set of forecast which solve the equations.

Notice that it would be very dangerous to estimate a system like this by ordinary regression. If SUPPLY did equal SEMAND in all historical years, then you would get exactly the same set of parameters (d,e,f) when you use regression on 5b as you did (a,b,c) when estimating 5a; you would not really have two different equations. Even if the equations were very slightly different, you could not rely on what you get when you subtract one from the other (as required in solving

them). In these kinds of situations, it is important to estimate the model as a system (Wonnacott and Wonnacott, 1977, chapter 22).

These situations arise in energy modeling, but the problem is usually not significant, mostly because we deal with dispersed system involving lagged responses. The systems estimation methods often lead to worse results, because e of their complexity, because the use of instrumental variables introduces random noise, and because of problems with "robustness" (discussed below). On the other hand, the simultaneity problem can be serious with fuels like LPG and other minor forms of oil, whose markets are very limited and respond quickly to price; our goal, in those cases, is to look for something like a "reduced form" model for each such fuel (e.g. in equations 5 first solve, to get log(Price)=g+h·LOG(GNP), and then estimate g and h).

Even the model in equations 5 still has one further simplification: all of the variables are assumed to be available for all historical years in you data base. (SAS will overlook a few missing values here and there, however.) It is possible to build econometric models which do not have this property, because they include "time-varying parameters" or "hidden variables;" however, this is not common at present, and the tools to estimate such models are hard to come by. One can work around this problem, to some degree; for example, if the population growth rate, "c" in equation 1, varies over time as a function of women's education (WED), then one might postulate that c=a+b·WED, and rewrite equation 1 as:

$$POP(t+1) = a \cdot POP(t) + b \cdot WED(t) \cdot POP(t) \quad (6)$$

Finally, for completeness, it should be emphasized that variables in an econometric model are not always simple time-series. Many authors will perform regressions on a data base of different observations at the same time, such as data from different states, and then use the results to predict the future. This is called forecasting based on cross-sectional analysis, and the results are usually unreliable at best, both in the short-term and in the long-term. For example, one of the first econometric equations ever studied was the classical consumption function:

$$C(t) = a + b \cdot Y(t) \quad (7)$$

where C is national consumption and Y is national income. In cross-sectional analysis, "a" was significantly larger than zero, and there seemed to be a large saturation effect in consumer spending. But in time-series, "a" was quite close to zero. For purposes of forecasting changes over time, the time-series version is the right one to use. In general, variations across space tend to be different from variations across time, and we have seen this lead to problems over and over again. (For example, see the discussion of "locational bias" in Werbos, 1983, Chapter 4.)

An ideal model should be able to account for variations over time and space both; however, without data from different times, it would be foolish to assume that one has an ideal model. Still, one can collect "pooled" data, which vary over time and space both, as we have often done (Werbos and Titus, 1978; Werbos, 1983). To use such data in packages like SAS can be slightly tricky, when you estimate a model containing lagged variables. In arranging our data (Werbos, 1983), we found it necessary to include a dummy year, 1973, to precede the years for which we had pooled data (1974-1981), and we inserted the SAS missing value code for all 1973 data. Observations 1 through 8 represented 1973 through 1981 in the first state, while 9 through 16 represented the second state, and so on. (Without this, the SAS "LAG" function would not have given us valid time lags.)

5. FIVE FUNDAMENTAL PRESCRIPTIONS

This section will provide a kind of back-door introduction to the theory underlying econometrics, by trying to explain five prescriptions for avoiding gross errors which are common even among professionalists.