

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

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 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:

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

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:

$$Y(t) = a \cdot Y(t-1) + b_1 \cdot X_1(t) + \dots + b_n \cdot X_n(t) + c \quad (2a)$$

$$Z(t) = c_1 \cdot Y(t) + c_2 \cdot Y(t-1) \quad (2b)$$

In equation 2a, Y(t) is the "dependent variable" -- the variable being predicted in that equation. Y(t-1) and X₁(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 b_n, 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.) X₁ 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:

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

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:

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

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

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:

$$\text{LOGQOIL} = \text{LOG(QOIL)}$$

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

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 from 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:

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

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

$$\text{SUPPLY} = \text{DEMAND},$$

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 just 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 DEMAND 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 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 $\log(\text{Price})=g+h \cdot \text{LOG}(\text{GNP})$, 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 $c=a+b \cdot \text{WED}$, and rewrite equation 1 as:

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

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.

- o Pay attention to uncertainty
- o Don't expect a free lunch when choosing specifications
- o Pay conscious attention to prior information
- o Don't expect to draw conclusions without adequate data
- o Check the historical track record of your model (This may be the most important)

Pay Attention to Uncertainty

None of the models above -- from equation 1 to equation 7-- say anything about uncertainty. They are all forecasting models, recipes for making base case projections. Even though there are many different schools of thought in statistics and econometrics, they all agree that uncertainty needs to be addressed explicitly as a central part of the analysis.

Broadly speaking, there are two major schools of thought here:

- o The purist school, which has done an admirable job of simplifying and unifying our understanding of statistical methods, and devising new and better and more elegant methods.
- o The utilitarian school, which has made life complicated and tricky all over again, by focusing on the intractable problems which occur in real-world forecasting. (This is quite different from the quick and dirty school, which pays more attention to deadlines than to quality problems either in theory or in the real world.)

Both schools have a great deal to contribute, but we incline towards the utilitarian school.

From the purist's point of view, regression simply cannot estimate equation 1 as it stands, as if it were a meaningful model of population growth. If you regress $POP(t+1)$ on $POP(t)$ with no constant term, then the model you are really estimating is:

$$POP(t+1) = c \cdot POP(t) + e(t) \quad (8)$$

where $e(t)$ represents a random disturbance, governed (generated) by a normal probability distribution. We sometimes call $e(t)$ "error," but statisticians like to think of it as something out there, in the real world, rather than an "error in the sense of "mistake.". Often we call $e(t)$ "white noise," to make this view explicit. Equation 8 is a "stochastic model," because the assumptions about the random disturbance have been made explicit as an integral part of the model.

When we look at the noise term explicitly, we can see immediately that there is something implausible about the model in equation 8. Equation 8 assumes that the noise comes from the same probability distribution in all years, implying that we should expect the same general size range for the noise in all years. If population grows by a factor of 10 in the period under study, this could be a very poor assumption about the noise; as a practical matter, this assumption would lead to an estimate of "c" dominated by the experience of the last few years, disregarding the earlier data. It is more plausible to expect that the noise will represent a certain percentage of the population, and that its size range will grow in proportion to the population, as in the model:

$$POP(t+1) = c \cdot POP(t) + e(t) \cdot POP(t) \quad (9)$$

In this equation, the overall noise terms -- $e(t) \cdot \text{POP}(t)$ -- grows in size in proportion to population; in other words, $e(t)$ -- which now represents noise as a fraction of the population -- still comes from a fixed probability distribution (imposing a fixed size range).

Equation 9 cannot be estimated directly in regression. However, now that we have a complete stochastic mode, it is legitimate to divide both sides by $\text{POP}(t)$, as we could with any algebraic equation; this yields the equivalent model:

$$\text{POP}(t+1)/\text{POP}(t) = c + e(t) \quad (10)$$

This can be estimated by regressing $\text{POP}(t+1)/\text{POP}(t)$ on no independent variables plus a constant term; in practice, this is just a matter of estimating the average value (mean) of $\text{POP}(t+1)/\text{POP}(t)$. It is more conventional, however, to use a similar but slightly more plausible alternative to equation 9:

$$\text{LOG}(\text{POP}(t+1)) = c' + \text{LOG}(\text{POP}(t)) + e(t) \quad (11)$$

which is equivalent to:

$$\text{LOG}(\text{POP}(t+1)/\text{POP}(t)) = c' + e(t) \quad (12)$$

More generally, equations like equation 8 -- which assume a constant size range for error when a constant size range is not plausible or does not fit the data -- are said to have a problem with "heteroscedasticity." This is a common problem, and algebraic transformations (like the above) are commonly used to overcome it. Sometimes, however, algebraic transformations are not a workable solution. For example, when the dependent variable is $\text{LOG}(\text{QOIL}/\text{QGAS})$, as in the standard "logit" specification for fuel choice, there is a heteroscedasticity problem which can only be resolved by resorting to weighted regression, which explicitly treats the size range of $e(t)$ as a function of other variables; the theory is given in Pindyck and Rubinfeld (1976), and applied in the PURHAPS model (Werbos, 1983, p.12,64). (This correction would have been desirable, but far less necessary, if we had worked with a simple time-series showing no order-of-magnitude variations in fuel shares.)

Besides heteroscedasticity, there are other possible problems with the theory that $e(t)$ is random and normal across time. For example, $e(t)$ may be correlated with its previous value, $e(t-1)$. When the standard Durbin-Watson test (available in SAS and other packages) gives a score much different from 2.0, it is conventional to use a different regression command -- regression with an autocorrelation correction -- to estimate the model under the assumption that $e(t) = r \cdot e(t-1) + a(t)$, where $a(t)$ is random; if r -- the "autocorrelation parameter" -- is not significantly different from zero, one can go back to using conventional regression.

Recently, many statisticians have begun to recommend a more careful study of the model residuals, $e(t)$, to see if they fit more complex "Box-Jenkins" models (Box and Jenkins, 1970). In theory, certain classes of Box-Jenkins models can represent the idea that forecast errors result from a combination of noise in the real world and noise in measuring what is happening in the real world. These kinds of models can reduce forecasting errors, but tests done for real-world multiple-equation models (Werbos, 1974; Werbos and Titus, 1978) suggest that it would be better to focus on the long-range track record of a model, as we will describe below. (Engineers have another way of estimating such stochastic models, but their formulation, unlike the statisticians' formulation, contains excess parameters and can almost never be uniquely estimated.)

All of these recommendations are based on the following fundamental theorem, an application of Bayes' Law, which underlies all inference from empirical data (in statistics or in other fields):

$$\text{Pr}(\text{Model}|\text{Data}) = \text{Pr}(\text{Data}|\text{Model}) \cdot \text{Pr}(\text{Model})/\text{Pr}(\text{Data}) \quad (13)$$

This states that the probability of a model being true, after we have observed a certain history of data, is the product of three terms. One of these terms -- Pr(Data) -- is the same for all models, and has no effect on our relative choice between models. Another -- Pr(Data/Model) -- refers to the probability that we would have observed what we did in the data if the model were true. For stochastic models, like equations 8 through 12, this term can be calculated directly by calculating what $e(t)$ would be in all the years of data (assuming a given estimate of the parameters as part of the "Model), and then using the normal probability distribution to calculate the associated probabilities. This term is called the likelihood function; it is a function of the parameter estimate, the model, and the data. The remaining term -- Pr(Model) -- represents the probability that a model would be true, a priori, before any statistical data are examined.

Most purists agree that it would be unscientific to account for Pr(Model) explicitly in statistical estimation. They argue that all possible models (and all possible sets of parameter estimates) should be treated as equally probable a priori. They argue that modelers should always estimate these models by finding parameter values which maximize the likelihood function. Most existing statistical packages do in fact maximize likelihood exactly (as in regression) or approximately (as in iterative methods which imitate the full information maximum likelihood command for estimating systems of equations).

Bayesian statisticians have argued that economists have important information, prior to statistical analysis, about the relative probability of different models and parameter values. We would agree, but would argue that the economists' information is very complex; it would be better to use the computer to produce a complete, graphic description of the likelihood functions -- the information found in the data -- and then count on the human being to account for his prior information after the statistical analysis is complete. This puts a heavy burden on the person doing the statistics, since it is not enough to just print our estimates of one final equation; it is essential to consider the range of uncertainty for all the parameter estimates, and to consider different ways of looking at the data.

Utilitarians (like us) go further, and argue that simple statistical models are never "true" in any absolute sense. They argue that your choice of estimation method should depend on the application of the estimates or forecasts. The overemphasis on definite, base case forecasts is a product of naive decision-makers, who have yet to understand well-known procedures for coping more honestly with uncertainty (Brown et al., 1974). Indeed, one may argue (Werbos, 1979) that probabilities, rather than expected outcomes, should be the main focus of long-range planning anyway; however, the efficient implementation of this principle involves many complexities (Werbos, 1987a; Werbos, 1986a). The utilitarian Raiffa has found that elite Americans tend to understate ranges of uncertainty by a factor of 3 or so, perhaps because they do not account for the limitations of the assumptions they use. This suggests a need for great care in using mathematical models built on expert judgement rather than empirical fact. Raiffa's followers, such as Rex Brown, have developed many techniques to train, improve and organize probability assessment by human judgement; nevertheless, the problem of bias remains difficult and fundamental. It is important that modelers help decision-makers think more clearly about alternative scenarios, rather than aggravate these biases. Even though it is very difficult to estimate probabilities objectively -- when technological and political forces are primary sources of uncertainty -- it should be possible to convey the nature of uncertainty in a useful way, and explain alternative viewpoints.

Utilitarians also tend to look for estimation methods which are likely to give more accurate forecasts even when it is hopeless to formulate a model which is "true" in an absolute sense; such methods are called "robust estimation methods." The problem of heteroscedasticity leads to a simple (though unconventional) example of robust estimation. Consider the simple model:

$$\text{ENERGY-USE}(s,t) = c \cdot \text{PRODUCTION}(s,t) + e(s,t) \quad (14)$$

where energy use is projected by state (s) and by the year (t). A purist would tell us to replace $e(s,t)$ by $e(s,t) \cdot \text{PRODUCTION}(s,t)$, and they divide by $\text{PRODUCTION}(s,t)$ to get a regression equation. If we do this, we are guaranteed that the percentage error in predicting energy use will average out to zero (i.e., positive and negative errors will balance out). If we had kept equation 14 as is, we would be guaranteed that the actual error in predicting energy use will average to zero; in other words, greater attention would be paid to bigger states. If the goal is to predict total energy use, then the latter is preferable; it would lead to more uncertainty in our estimate of c, in theory, but it would also guarantee that we are estimating that version of c which is right for our application. If we admit that equation 14 is only a simplification, then we have to accept that the version of c which minimizes one error measure will be different from the version which minimizes another. In technical terms, there is a tradeoff here between statistical efficiency (i.e., random uncertainty in our estimate of c) and statistical consistency (estimating the right c). Tradeoffs of this sort are quite common, and often require some sort of ad hoc compromise.

Don't Expect a Free Lunch When Choosing Specifications

Choosing the equations of a model is a difficult process, whether the model is econometric, judgmental, or engineering-based. The process is essentially the same for all three, except that econometricians normally restrict themselves to using variables for which they have data. Econometricians often start from a general theoretical model and translate it into its implications for observable variables; there is no need to represent the entire mechanism by which variable A affects variable B if the ultimate impact is represented correctly. Also, when doing econometrics, you usually consider several alternatives, and use empirical results to decide which version to select in the end. In fact, you typically try out new alternatives after you have studied the results and looked for explanations of what is going on.

There are some analysts who offer you a hope of forecasting without resort to this difficult process. They often suggest that "simple time-series analysis" or "simple econometrics" can be an alternative to the labor and uncertainty which comes with explicit models. In actuality, this is an illusion (though the explicit models of econometrics are simpler than most engineering models). For example, "simple Box-Jenkins analysis" (Box and Jenkins, 1970) offers more complicated models of noise than regression assumes; it essentially offers yet another complex correction to explicit models (Werbos, 1974). The vendors of "simple" analysis typically apply statistical methods to a simple forecasting model, such as:

$$Y(t+1) = a + b \cdot Y(t), \quad (15)$$

where Y is the variable you are trying to forecast. Admittedly, this model is sometimes worthwhile. Admittedly, simple models in general tend to be more robust than complex models, ceteris paribus. Some salesmen have suggested that this approach can be applied to electricity demand, to save utility planners from the pain of using models which require forecasts of local industrial growth, which are fraught with uncertainty. However, this economic uncertainty is real, and unavoidable; a forecaster can hide the uncertainty from his clients (which does them a disservice), but the economic uncertainties are there and will affect electricity demand. If economic growth is known to be central to electricity demand, then it should be reflected in the model. In general, the choice of a model should be based on a careful analysis of what is known about the variable being modeled, and what is shown in the data; there is no magical way to escape this process.

Forecasting problems in private industry are sometimes so complex that analysts cannot devise an adequate specification, even when data are plentiful. In such situations, a full-scale "neuron network" system may be useful. The best neuron network systems (Werbos, 1987a; Werbos, 1988; Werbos, 1989) are essentially equivalent to a massive automated search through all possible specifications -- linear and nonlinear -- to find that specification which minimizes some combination of forecast error and model complexity. The prior knowledge of the analyst

is not used at all (except in the construction of the data base.) At EIA, we have yet to encounter such situations.

In one recent situation (IFS, 1986), we had access to a massive data base on fuel-switching in which we didn't know what to expect. Elaborate cross-tabulations in SAS were very useful in helping us form hypotheses, and then formulate and estimate econometric models.

All of these examples emphasize that a modeler should take the time to devise specifications carefully. Some students are willing to do this, but expect to be given exact rules on what kinds of specifications to use. Once again, they are looking for a kind of free lunch, and they can find a few misleading papers in journals which give them the rules they are looking for. In actuality, the choice of specification should be based on a translation of your prior knowledge (Pr(Model) in equation 13) into mathematical equations; there is no set rule for what the equations must look like, but there are guidelines for how to do the translation. A good econometrician should have some familiarity with the guidelines for translation (Brown et al., 1974; Forrester, 1961) which have been developed for models in general. Econometricians have developed further guidelines, but they are too numerous to cover here; still, please do consider what happens to your model when the independent variables take on extreme values, and do consider whether the forecasts would change the way you want them to in response to a small change in the inputs (as a function of other inputs). Also consider whether the specification really could represent alternative points of view (e.g. large and small price elasticities) through different parameter estimates.

This notion of translation between human knowledge and mathematics is so vital that it merits several examples.

First of all, translation from English into mathematics may be compared with translation from Chinese into English. In Chinese, one can make statements like "man see horse." In English, this could mean that "a man saw a horse", or that "every man sees a horse sometime in his life", or that "those three men are looking at a horse", or that "this man will see a lot of horses", etc. In order to translate from Chinese into English, one has to decide what tense to give the verb "see", what number or article to put before the word "man", etc. A good translator will make these decisions based on a careful understanding of the context in which the statement appears. Even then, several interpretations may still make sense; in that case, the translator may go back to the author of the Chinese statement, and ask which alternative would be used. Note that the translator can state the alternatives to someone who only speaks Chinese; the Chinese language permits ambiguity, but does not require it.

An irresponsible translator would not try to understand the Chinese original; instead, he would follow a mechanical rule, such as assuming the present tense in every sentence which does not explicitly refer to the future or the past. Irresponsible translators can easily produce paragraphs in English which look downright silly (as in the instruction manuals which come with certain imported products). In translating from English into mathematics, one can produce silly mathematics just as easily, if one is not careful about the role of time in the equations.

Second, consider a question which the Energy Modeling Forum brought up in 1985; "How much fuel-switching has there been between oil and gas in response to prices in manufacturing?" Two modelers came up with completely different answers to this English-language question, based on the same set of data (Annual Survey of Manufactures) at the State level from 1974 to 1981. One modeler (David Reister of Oak Ridge) translated the English-language question as follows: "In any given year, was the market share of natural gas in manufacturing as a whole much greater in those States where the gas price was a smaller fraction of the oil price?" The other (myself) translated it as: "In any given industry, was the change in market share from one year to the next much greater in those States and years where the change in the price ratio was also great?" These are two different questions, and it is not surprising that they yield different

answers. At EIA, we have tested both kinds of specifications, and were not surprised that the latter type led to much smaller forecasting errors.

Translation back from the statistics into intuitive terms is just as important. For example, a few years ago we reviewed a major paper on new car Miles per gallon (mpg), which developed a model which predicted that a doubling in gasoline prices would double mpg in the future. Working through their vehicle attribute equation, we discovered that this forecast depended on the assumption that new cars would be eight feet tall or eight feet long in order to get higher mpg.

Pay Conscious Attention to Prior Information

Often forecasts try to provide a "most likely" view of the future, based on all of the information available. Historical information is only part of that information. Expert judgement, private sector plans, and engineering information are also part of that information. In an ideal world, we should not have to choose between econometric forecasts versus engineering forecasts, and so on; we should develop forecasts which have the highest probability of coming true, conditional upon all three kinds of information. In an econometric model, this can be approximated by choosing specifications and altering parameters, where necessary, to reflect such information.

This kind of adjustment is a tricky process. There is a risk of confusing the final user, who may not be able to tell what comes from historical trends and what comes from adjustment. Also, when adjustments are made on the basis of judgments, political biases and wishful thinking easily enter in, and cause further confusion. Adjustments based on population wisdom which in turn depends on past history may represent a "double-counting" of history or far worse. Therefore, there is much value in having some forecasters -- such as academics -- produce pure econometric models and leave the discussion of other factors to their verbal discussion sections.

The Energy Modeling Forum (Werbos, 1987b) has given us some examples of how this kind of adjustment can be done and explained to the reader. For example, if an historical trend (reported, say, as $c \cdot \text{year}$ in an equation to predict energy intensity) can be clearly explained as the result of using a single technology throughout the historical period, and if we are quite sure that a radically different technology will come on and dominate the forecast period, then an econometric model should be adjusted to reflect our best knowledge of the new technology. Conversely, if new technologies are expected in the future, but are numerous and hard to predict exactly, and if there were also new technologies coming on line in the historical period, one is better off trusting the econometric model.

Don't Expect to Draw Conclusions Without Adequate Data

Econometric models, when estimated, make a statement about cause and effect relations. For example, in equation 6 above, if "b" were estimated as a negative number, this would say that increases in women's education can reduce population growth by some amount. If b were estimated as a positive number, it would say the reverse. In either case, the estimated value may be a fluke, a coincidence due to a relative shortage of data. To see if this is likely, we need to examine the "standard error" of b, which is possibly the most important statistic printed out by standard regression programs. By and large (Wonnacott and Wonnacott, 1977), the true value of b will be equal to the estimated value plus or minus the standard error, in seventy percent of the cases; it will lie within two standard deviations in ninety-five percent of the cases. Old-fashioned statisticians would say that b "is not significantly different from zero" when the value $b=0$ was between these confidence limits; however, it is better simply to report what the standard errors are, to make it clear to the reader how big b still might be (given the limitations of the data).

Large standard errors may result from any of the following:

- o Lack of data. (Having four times as many independent observations cuts the standard errors in half. However, with pooled data, the observations -- e.g., from neighboring states -- may not be entirely independent, and the true standard errors may be larger than the reported ones.)
- o Lack of an adequate specification. (Cutting historical error in half cuts standard errors in half.)
- o Correlations between the independent variables, or "multicollinearity." (This represents a qualitative lack of data -- a lack of data on situations where the independent values have different values.)

Many social scientists do not appear to realize that standard errors really do account for the effects of multicollinearity. When two independent variables do correlate very strongly with each other, there is no magic procedure to solve the problem; an honest statistical analysis will simply report that there is not enough data available to decide which variable has an impact on the dependent variable. When there is strong multicollinearity (as hinted at by large standard errors), but no really strong correlation between two variables, one should suspect three-way patterns of correlation; to locate these kinds of patterns, one can perform an eigenvector analysis of the correlation matrix, and look for the eigenvector whose eigenvalue is closest to zero. (Belsley, Kuh and Welsch have discussed these kinds of diagnostics.) When there are many variables involved, and when forecasts will be made for situations where the independent variables continue to correlate with each other, methods like "ridge regression" may be better than ordinary regression when multicollinearity is suspected (Dempster et al., 1977). When a model must be estimated, but the data are inadequate, one must generally fall back on prior information (and flag the resulting uncertainty).

Check the Historical Track Record Of Your Model

There are many economists who test out alternative specifications, and publish whatever gets the highest "R-squared" score as printed out by SAS. This often leads to disaster, because R-squared scores are not comparable between equations which represent the dependent variable in different ways; for example, an equation which predicts energy per unit of output will often be more accurate than an equation which simply predicts energy, but will often have a lower R-squared score (because the dependent variable has less variance). The situation is even worse with complex statistical methods as used in fields like cost function estimation; there, the "adjusted R squares" are often aggregate constructs whose relation to forecast error may be quite tenuous.

As a first step, one can try to compare mean square error (MSE) across models, because it is reported by SAS and is more often comparable between equations. As a second step, one can simply use the alternative equations to predict the same basic variable (e.g., energy consumption), and calculate the average error; this can be done in SAS by using the "OUTPUT" option to output the regression predictions to a file, and by using the numerous SAS utilities to calculate the implied predictions of energy and their errors.

In practice, there is no substitute for trying to understand what is in the data, as directly as possible. Predictions and actual values should be plotted against time, where possible, and the differences explained. This provides a basis for going back and changing the model (or better understanding its weaknesses). Plots like these are important both in estimating a model and in explaining the model to others. Tukey of Princeton has written a book on Exploratory Data Analysis, describing additional techniques for better understanding the residuals graphically.

In the past, some econometricians have routinely used "dummy variables" (1 in some years and 0 in others) or other procedures to throw out "outliers," observations which are hard to explain using their forecasting model. (Some statisticians have also recommended maximizing the

1.5 power of error instead of the square error, which has much the same effect.) More recent authors, like Belsley, Kuh and Welsch, have stressed to need to study the outliers (and other "influential observations") rather than simply throw them out, because they may be crucial to what your model is trying to forecast and may be important as a guide to a better model. For example, the oil shortages of 1974 and 1979 were "outliers." but a model which ignores them is a poor guide to reality. Again, the hard-to-explain observations should be obvious in a plot of predictions and actuals, but more sophisticated tools exist for identifying them.

In practice, we have also found that there is no substitute for performing a "dynamic simulation" test of a model, if you are considering the use of a model which contains a lagged endogenous variable. This is quite different from evaluating the "predicted values" which come out of a standard regression command. For example, if you were estimating equation 1 over the period from 1967 to 1985, the "predicted population" for 1980 would be calculated as "c" multiplied by the actual population in 1979, in a regression package. In dynamic simulation, the prediction for 1980 is calculated as "c" multiplied by the prediction for 1979, which in turn is calculated as "c" times the prediction for 1978, and so on. The regression test would be appropriate, in theory, if predictions one year ahead were all that you care about. The dynamic simulation test would be better if you planned to forecast further out into the future, or if the real concern for policy is the eventual result several years into the future.

A purist would argue very strongly that the regression test is adequate, if one has faith in the truth of one's model. He would argue that those lacking in faith should look for better models. A utilitarian would argue that all models are oversimplifications, and that faith without tests is no way to do modeling. Experience has shown that cumulative error tends to be quite important (or even overwhelming) in models containing lagged endogenous variables, regardless of their theoretical virtues. More to the point, it has shown that such errors can be avoided either by specifications without lagged endogenous variables (if such can be found, with otherwise comparable MSE scores) or by a new form of robust estimation.

An example of the former comes from our PURHAPS model: by filtering the effect of energy prices over time, we can represent the notion of capital-embodied price responses just as effectively as do recent academic models based on lagged dependent variables (previous period energy intensity); however, our older version may be more robust. On the other hand, there are many models (especially models which assess changes induced by policy) which have much higher error levels and much crazier parameter estimates when lagged endogenous variables are not included.

To estimate models with lagged endogenous variables, several authors -- including Larry Klein of Wharton, and myself (Werbos, 1974) independently -- suggested several years ago that models could be estimated by directly picking parameters so as to minimize errors in dynamic simulation. This could be implemented in practice by doing the dynamic simulations on a PC package like Lotus, and adjusting the parameters by hand to minimize the error in dynamic simulation. This form of robust estimation may be somewhat extreme; however, we have found (Werbos and Titus, 1978; and Chapter 4 of Werbos, 1983) that it is possible to compromise between this approach and regression, and still allow for a noise term in the model (allowing for uncertainty). Dynamic robust estimation methods of this sort have cut errors in half in a number of applications (where lagged endogenous variables were important), and have even done better in short-term forecasting (Werbos, 1974; Werbos and Titus, 1978). The compromise method looks similar to "exponential smoothing" methods which are essentially equivalent to the Box-Jenkins methods discussed above; however, they weight the square error in different ways, and this difference in weighing will probably be the key feature even of more advanced methods along the same lines.

When calculating error for a system of equations, one may simply use a weighted sum of the error for different variables (including "filtered" variables). A utilitarian would argue strongly for doing this, and for weighing each dependent variable's error according to the importance of that

