October 20, 1978

THE EVALUATION OF ECONOMETRIC MODELS AND MODEL CHANGES

Scott Goldsmith

This paper reviews and discusses the generally accepted criteria for the evaluation of econometric models and shows how one might use them in the context of a potential model change.

I. The Starting Point: Initial Model Design

The starting point for every model to be used in economic analysis is economic theory. The theory tells the researcher what the general form of the relationships between the various economic variables should be. Without a specific theory, the researcher does not know what his data, equations, and model are telling him.

Consider the following equation:

$$Q = f(P,X)$$

where Q = quantity of a good
P = price of that good
X = a list of other related variables

Is this a demand equation or a supply equation? It could be either in this form. A theory is necessary to choose the list of variables, X, to determine whether it is a supply or demand equation.

For the policy analysis purposes for which most econometric models are designed, a complex model is often preferred to a simpler but naive model. This complexity allows the model to be used to do a great variety of analyses. However, as soon as the researcher starts to make his model complex, he begins to encounter problems which require that he balance off different objectives. Most of this paper will discuss these tradeoffs, but a simple example will indicate the principle involved.

Consider two models: one, the DRI model of the U.S. economy or some equivalently large and complex econometric model and, the other, the following "simple" model.

$$GNP = a + bt + cT$$

where GNP = U.S. gross national product
t = time
T = income tax rate

Which model is preferred? Both are designed to calculate U.S. GNP and can accomplish policy analysis in which the effect on GNP of changing the income tax rate is determined. To make the question interesting, assume that the simple model predicts GNP better than the complex model over some period of time. This could actually be the case.

The complex model allows the researcher to analyze a large variety of tax changes and to look at the impact of the tax change on a large number of variables not present in the simple model. In addition, the complex model is based on economic theory rather than making GNP a function of time, which may improve its capability of projecting changes in the relationship of GNP to time. The choice then is between accuracy in one task on the one hand, and flexibility in use on the other. The answer depends on expected uses of the model which determines which criteria is more important, as well as the researcher's budget.

II. Data

After having developed a framework for the model based upon economic theory and the use to which it will be put, the researcher attempts to collect the data necessary to develop the parameters which define the model relationships. At this point, many models fail because data in

the real world never conforms to the expectations of the theoretical economist. Unlike the physical sciences where the researcher creates his own data through experimentation, the economist must generally rely on data collected by someone else for some other purpose. For this, however, he considers himself lucky, because for many of the variables in his theoretical model, there is no data at all.

Just because data is available on a variable is no guarantee that it is usable. In Alaska, in particular, all data should be presumed guilty of substantial "noise" (variation of reported values from the true values) until proved otherwise. The difficult problem is to determine under what conditions data with noise in it should be used anyway, and under what conditions the variable should be omitted from the analysis altogether because, on balance, it will increase the noise of the whole model. There are few easy choices in econometric model building.

Occasionally, one can clean up a data series or artificially create a data series using partial data like census information. This is a valuable technique for allowing more complex models to be built, but the same caveat about noise and noise creation is applicable here.

III. Criteria for Model Evaluation

III.A. Statistical tests

A standard set of statistical tests is available to determine, in a statistical sense, whether an individual equation in a model should be judged to be reasonable. R^2 increases as the variation of the dependent variable is better "explained" by the independent variables. The t value of individual coefficients is larger if that particular variable significantly "explains" any of the variation of the dependent variable. The list goes on. Other things being equal, high R^2 and t values are preferred. Unfortunately, these tests cannot tell the

researcher whether or not the structure of the equation corresponds to theory, however, or that it will predict and otherwise perform well. It just tells one how the equation "fits" the data. "Explain" used above really means "fit," and it should be noted that time series data, the kind more often used by economists, generally "fits" well.

If several equations in a model have been estimated simultaneously using a more exotic technique such as two-stage least squares, the same tests generally apply and the same degree of confidence should be engendered in the researcher by good statistical test results or good "fit."

Even confining the choice of equation to statistical criteria does not obviate the need for the researcher to use his judgment. One form of the equation may yield a high R^2 while the alternative, very significant t statistics. The choice will partially depend on the use of the equation. For forecasting purposes, R^2 is important; while for hypothesis testing or measurement of elasticities, large t statistics are preferable.

III.b. Formal simulation tests

For a single equation model, one can calculate the R^2, t, and the error of forecast as indications of the goodness of fit of an equation and its ability to recreate, through historical simulation, the actual historic values. A historical simulation involves running the simulation model over the historic period to see how well it "tracks" the variables it is predicting.

With a multiequation model, the situation is different if there is any interaction among the variables (and it would be a dull model if there were none), because in such a case, the statistics of all the individual equations may indicate a good "fit" and yet upon running a simulation, the model could get way "off track." This is because the interactions of the model mean that the simulation value for a particular variable now could be determined by several equations interacting simultaneously.

Table 1. MAPE COMPARISON

	MAP 65-76	Phil I	Phil III	Phil IV	Northeast Corridor	Buffalo	Los Angeles	Mississippi
gross output	.78	4.32	.97	.98	2.05	1.87	2.08	.94
total employment	1.96	1.56	1.21	.66	1.40	3.39	.88	.41
personal income	3.72	6.69	1.50	1.55	3.13	8.42	1.45	.70
population	3.15	1.02	1.49	1.27	.78	2.50	NA	NA

Source: Norman Glickman, "Son of 'The Specification of Regional Econometric Models'." Papers of the Regional Science Association, Volume 32, p. 165.

In such cases, high R^2 values are of little comfort. Several measures have been developed to measure how well a model of many variables is able to replicate the historic period by taking various aggregations of the individual differences between actual and estimated variable values. Among these measures is the mean absolute percent error or MAPE. This measure calculates for each variable the mean of the absolute percent differences between the actual and predicted values. One can thus evaluate the MAPE for a single variable (as opposed to equation) or for all variables together.

Unfortunately, there is no specific guideline to tell the researcher when a satisfactory MAPE value has been obtained either for a single variable or for the complete model. One can only compare MAPE values between model runs and between models. MAPE values for important variables of the MAP model are shown in Table 1 and can be put in perspective by comparing them to other representative regional econometric models.

The difficulty with a complex model, because of equation interactions, is that the variables are not independent of one another. Thus, the attempt to improve the MAPE statistic for one variable through a model change will normally change the MAPE statistics for several other variables. One must be resigned to the fact that it is not possible to simultaneously improve (reduce) all MAPE statistics, so the final form of the model chosen should be the one that "tracks" for the most important variables.

If the MAPE statistics are generally bad, it means that the model is not simulating well; and in order to improve on its simulating capabilities, it will probably be necessary to sacrifice some of the equations with good statistics ("fit" well according to R^2, etc.) in favor of equations that have less desirable R^2 but which nonetheless improve the ability of the model to simulate. It is the rare model builder who has not been forced into this tradeoff and, other things being equal, it is the more

simple-minded model in terms of structural interrelationships which will have the better MAPE statistics.

III.c. Theoretical reasonableness

A model may consist of equations which all have desirable statistical properties and yield low MAPE values for all variables when run over the historic period. In spite of this, it may yield "unreasonable" results when employed to do a future projection. The test of reasonableness is not quantifiable but rests upon the economic theory which was the starting point for model construction.

For example, if a regional econometric model is constructed on the basis of a neoclassical theory of factor movements, it is to be expected that over time the factor returns in the region would move toward national averages. Thus, if the model simulated wage rates growing away from parity with national averages, this would be a signal to the researcher that the model was not operating consistently with the theory, and adjustments should be made.

Alternatively, simulation results may not be what one expects, or "counterintuitive," in spite of seeming consistency with the theory. Such a result, of course, requires close investigation of the equations determining the result. If intuition is to be the final arbiter, then, of course, there is no reason to build a model at all. It is a strength of a complex model that it is able to keep track of relationships which an individual, relying on intuition and a limited memory bank, could easily overlook or lose track of.

Several types of tests of model theoretical reasonableness are common. The model should be able to pick up turning points in the time path of an important variable or significant changes in the growth rate. For the Alaskan economy, the Alyeska years would be such a case. The model should give dynamic responses to large changes in exogenous

variables and policy stimuli that are consistent with theory and recent experience. Ratios between important model variables will be (theoretically) constrained to remain within particular ranges. Monitoring these ratios is an additional check on reasonableness. For example, the ratio of employment to population could not be reasonably expected to approach zero or one.

III.d. Sensitivity tests

A well-designed model should be insensitive to a variety of changes. These include the time period during which a simulation is initiated, small changes in the values of individual coefficients, and small changes in the time paths of exogenous variables. As with theoretical reasonableness, there are no established standards of sensitivity by which the researcher can judge his results.

III.e. Stability

In a simple linear model, one can test for stability by solving the characteristic equations for the roots. Stability means simply that over time the variables in the model will come to rest at an equilibrium position if they are moved off an equilibrium position by an exogenous shock. In an unstable model, the variables will forever move away from an equilibrium if once disturbed.

Large complex models are more likely to be unstable because of the interaction of the equations in the determination of values for the variables. Unfortunately, for large models which are usually nonlinear, one cannot simply solve the characteristic equations. Stability can only be determined by "putting the model through its paces" in a series of simulations covering as long a period of time as feasible and covering as much variation in the exogenous variable values as plausible. In this way, one develops a "sense" of the model's stability properties.

A little bit of instability is not necessarily a bad property for a model to have. The real world, after all, is not always stable and does not always conform to the comparative static framework of the neoclassical economist. Exogenous changes to the system do not occur sequentially nor are factors, technology, tastes, etc., constant over time.

IV. Model Updates and Changes

Model updating is carried out any time that new data points or new variables become available or when economic theory and model results suggest that the present formulation of the relationship between the variables is unreasonable. As with model development, there are several criteria for judging the value of a potential model change. One weighs the criteria on the basis of what the model will be used for. The choice will probably result in a tradeoff among the criteria. This is the sense in which econometric modeling is an art.

In a model that is being used primarily for policy analysis rather than prediction, the t statistics are more important than R^2 and theoretical reasonableness is more important than low MAPE values. In predictive models, the opposite would probably hold true.

* * * * *

Several important observations derive from this discussion:

1. The more complex a model is in terms of variable interrelationships, the more likely it is that low R^2 will be encountered with some equations, MAPE values may be poor, and instability becomes a potential problem. Generally, the richness of model detail outweighs these concerns.

2. In a complex model, an individual equation must be analyzed in the context of the whole model rather than "out of context."

3. Because of multiple criteria for determining model reasonableness, each modeler assigns his own weights to the individual criteria in model construction. Thus, each model bears the imprint of the designer.

4. Choosing a structural form for an equation or deciding whether to include or exclude a specific variable in a particular equation is never a straightforward matter.

5. There is no perfect model which absolutely satisfies all criteria. All can be criticized and improved upon. But because no model completely satisfies all criteria does not imply the models are not useful.