ALTERNATIVE TESTS FOR THE SELECTION OF MODEL VARIABLES

Nathaniel J. Mass

Peter M. Senge

System Dynamics Group
Alfred P. Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

This paper contrasts two approaches to testing the importance of model variables: single-equation statistical tests and model-behavior tests. The paper demonstrates that, both theoretically and operationally, tests which analyze the impact of individual variables on model behavior are better suited to the task of selecting model variables. Conversely, the statistical tests should not be viewed as tests of model specification per se, but as tests of a particular type of data usefulness. When viewed as tests of data usefulness, the statistical tests have a clear, albeit quite narrow, role in model validation: they warn the modeler when available data do not permit accurate estimation of a model parameter. However, as a detailed example illustrates, a model relationship may be difficult to estimate yet extremely important for overall model behavior.

TABLE OF CONTENTS

I. INTRODUCTION

One of the most difficult and subtle tasks confronting the mathematical model-builder is selecting appropriate variables and functional relationships. In specifying each equation, the modeler faces two distinct problems. First, one must posit an a priori hypothesis regarding causes of change in the dependent variable of the equation. The a priori hypothesis, which may be based on direct observation, prior theory, or both, identifies the variables believed to be significant determinants of change in the dependent variable. The a priori hypothesis also specifies how these determinants are to be combined. Second, the modeler must have some means of testing whether or not, given the available empirical information, the variables and relationships obtained from a priori reasoning are in fact important. Based on the results of such testing, the modeler may reject certain variables as relatively unimportant, and may thereby begin to refine initial causal hypotheses.

This paper addresses the second aspect of the problem of variable selection--testing the importance of model variables. The paper contrasts two approaches to model testing--single-equation statistical tests and model behavior tests. Single-equation statistical tests are tests which compare an individual model equation to statistical data. Two such tests are the popular t-test of parameter "significance" and the partial correlation coefficient, both of which focus on the impact of one particular "explanatory" variable on a "dependent" variable. Model behavior tests measure the importance of an hypothesized impact on one variable on another for the behavior of a complete system model. The key distinction between the two testing approaches, then, lies

in whether or not a test examines closed-loop feedback response to altering the relationship in question.[1]

The major theme of the present paper is that only tests that examine a variable's influence on overall model behavior provide a sound basis for assessing an individual variable's importance. On the other hand, the single-equation statistical tests employed extensively in modeling practice can be extremely misleading if used as the sole guide to rejecting or accepting a variable in a causal model. Both theoretical and practical arguments are presented in support of the superiority of model behavior tests.

The arguments presented below pertain especially to the social sciences, where alternative theories frequently match the statistical evidence equally well. Whereas most social scientists attribute the presence of alternative "equally valid" theories to the paucity of reliable data, the present paper suggests that the operating philosophy of theory testing may be at least equally at fault.

Criticism of the single-equation statistical testing viewpoint can be traced back to John Maynard Keynes who, in reviewing Jan Tinbergen's Statistical Testing of Business Cycle Theories: A Method and Its Application to

---

[1] Most statistical tests employed in econometric practice belong to the single-equation class, regardless of whether the statistical test is based on a recursive, simultaneous, or even multiple equation (e.g., Zellner [28]) estimator. On the other hand, statistical tests using an estimator based on the Kalman filter (see Kalman [15], Schweppe [25], or Peterson [21]) belong to the class of model behavior tests. To see the distinction, consider that the former set of estimators do not involve simulating (or, in the case of an extremely simple system, analytically solving for) model behavior in computing a likelihood function, while a Kalman-filter-based estimator does.

Investment Activity, argued that

> The method [multiple-correlation analysis] is one neither of discovery nor of criticism. It is a means for giving quantitative precision to what, in qualitative terms, we know already as the result of a complete theoretical analysis. . . . How far are these curves and equations means to be no more than a piece of historical curve-fitting and description, and how far do they make inductive claims with reference to the future as well as the past? If the method [multiple-correlation analysis] cannot prove or disprove a theory, and if it cannot give a quantitative guide to the future, is it worthwhile?[2]

More recently, many social scientists have once again begun to question the power of well-established statistical tests (see, for example, Morrison and Henkel [2]). Therefore, the present paper can be seen as part of a methodological debate which has persisted for over thirty years. Within this context, the paper is, however, unique in that it prescribes a concrete and broadly applicable alternative to single-equation statistical testing.

II. STATISTICAL TESTS FOR THE SELECTION OF MODEL VARIABLES

This section discusses the type of single-equation statistical tests employed frequently in modeling practice as a guide in selecting model variables. The section shows that such tests actually measure the degree to which available data permit accurate estimation of model parameters and, thus, should be viewed more as tests of data usefulness than as tests of model specification. A case study in a subsequent section illustrates how, if used

[2]Keynes, [16], pp. 567, 569.

as a basis for selecting model variables, the statistical tests can seriously mislead the model-builder. The discussion focuses on two particular statistical measures--the t-statistic and the partial correlation coefficient.

A. The t-Test[3]

The t-test provides the modeler with a measure of the confidence he can place in an estimated parameter value. A statistical parameter estimator is a random variable.[4] The estimator is a random variable because the equation being estimated, and therefore the data upon which the estimate is based, is assumed to have a random component. Because it is a random variable, the estimator has a mean and a variance. If the estimator has a large variance, little can be said with confidence about its accuracy. That is, even if the mean of the estimator equals the true value of the parameter being estimated (that is, the estimator is "unbiased"), a large variance means that the probability that the parameter estimate is close to the true parameter value is low. The t-statistic provides a measure, based upon a formal statistical hypothesis test, to determine whether or not the variance of the parameter estimator is "too large."

[3]The following exposition of the t-test does not attempt to explain fully the mechanics of the test. For such explanation, the reader should refer to an introductory econometrics text (for example, Theil [28]).

[4]The term "estimator" refers to the estimation technique as an operator which converts a set of data into a set of parameter estimates. The distinction between the estimator and the estimate is analogous to the distinction between a random variable and one particular value of the random variable.

In a linear regression, the t-statistic is computed as the ratio of the computed parameter estimate β to the estimated standard deviation of the parameter estimate $\hat{\sigma}_{\hat{\beta}}$:

$$\text{t-stat} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \quad . \qquad (1)$$

As the estimated standard deviation $\hat{\sigma}_{\hat{\beta}}$ increases relative to the parameter estimate $\hat{\beta}$, the t-statistic diminishes. Given a certain set of assumptions (the realism of which is not in question here[5]), the t-statistic becomes a tool of statistical inference. For example, a t-statistic whose absolute value is greater than 2.3 permits the modeler to <u>infer</u> with a probability above 0.99 that the true parameter is non-zero.

As a measure of confidence in a parameter estimate, the t-statistic guides the modeler toward judicious use of the available sample of data. Passing the test tells the modeler that an estimate is fairly "tight"--that is, the estimated standard deviation of the estimator is small relative to the estimated parameter value. Failing the t-test tells the modeler that the estimated standard deviation of the estimate is too large relative to the estimated parameter value. However, neither outcome tells the modeler whether the underlying parameter β or the associated variable is in any sense "important" for the model being estimated. As will be illustrated, the statistical significance

[5] The assumptions underlying the t-test include perfect specifications of the model being estimated (including zero-mean, normally-distributed noise inputs in each equation) and perfect measurement of all variables. Senge [26] discusses the realism of these assumptions and how typical violations of the assumptions affect the accuracy of least-squares parameter estimates.

or "tightness" of an individual parameter estimate and the importance of the associated variable may differ markedly.[6]

Therefore, the t-test is best interpreted as a test of a particular type of data usefulness, not as a test of model specifications. Failure to pass the t-test means that the available data do not permit accurate estimation of the parameter β. Conversely, passing the t-test means that the data are useful for the purpose of estimating β--that is, β can be estimated with suitable precision. Although such tests of data usefulness may play an important role in the overall validation process, they must be sharply distinguished from tests of model specification.

B. The Partial Correlation Coefficient

The partial correlation coefficient provides the modeler with a measure of the incremental contribution of a single right-hand-side ("explanatory") variable in accounting for variation in a dependent variable. Denote the explanatory variable in question x_h. If one computed the fit of the estimated equation twice, once with the variable x_h included in the equation and once with x_h excluded, the partial correlation coefficient r_h could be determined from the two computations of the coefficient of multiple determination R^2 (R_h^2 corresponds to the case when x_h is omitted from the equation):

[6] For example, multicollinearity and measurement error are two common causes of statistical insignificance. Neither multicollinearity, which arises when one variable on the right-hand side of a regression is highly correlated with another right-hand side variable, or with a linear combination of right-hand side variables, nor measurement error necessarily imply that the model itself is defective as a causal description of the real system or that individual model variables are "unimportant."

$$r_h^2 = \frac{R^2_{total} - R^2_h}{1 - R^2_h} \qquad (2)$$

An econometrician would say that the partial correlation coefficient measures the "amount of variance explained" by the variable x_h. Although intuitively appealing as a test of the importance of the variable, the partial correlation coefficient actually yields no information not already provided by the t-statistic, as can be seen by the following relationship between the two measures (t_h denotes the t-statistic for the coefficient associated with the variable x_h).[7]

$$r_h = \frac{t_h}{\sqrt{t_h^2 + n - k}} \qquad (3)$$

where r_h = partial correlation coefficient for x_h

t_h = t-statistic for x_h

n = periods of data available

k = number of coefficients to be estimated in equation.

Taking the partial derivative of r_h with respect to t_h,

$$\frac{\partial r_h}{\partial t_h} = \frac{(n - k)}{(t_h^2 + n - k)\sqrt{(t_h^2 + n - k)}} \quad . \qquad (4)$$

Equation (3) implies that, as the t-statistic approaches zero, the partial correlation coefficient likewise approaches zero. Equation (4) shows that the partial correlation coefficient decreases whenever the absolute value of the t-statistic decreases ($\frac{\partial r_h}{\partial t_h} > 0$), provided there are more data points

[7] Theil [28], p. 174, provides a derivation of Equation (3).

than parameters to be estimated ($n - k > 0$).

Because the partial correlation coefficient is so closely coupled to the t-statistic, it may indicate that a variable contributes little in "explaining movements" in a particular dependent variable when, in fact, the relationship between the two variables is simply difficult to measure given available data. Therefore, the partial correlation coefficient is clearly not a reliable guide to model specification. To obtain a more reliable measure of the contribution of variable x_h in explaining observed movements in a dependent variable, it is necessary to analyze the behavior of the system of feedback relationships within which the relationship in question is embedded.

III. MODEL-BEHAVIOR TESTS FOR THE SELECTION OF MODEL VARIABLES

Given that single-equation statistical measures, such as the t-statistic and the partial correlation coefficient, do not provide reliable measures of the relative importance of different variables, what alternative techniques might provide insight into this critical problem? Section III attempts to provide some guidelines and direction for addressing the importance of the hypothesized impact of one variable on another. In particular, the discussion focuses on an approach to testing which assesses a variable's influence on model behavior.

The model-behavior testing approach entails three principal steps. Suppose the modeler seeks to determine whether or not variable X is an important determinant of observed oscillations in variable Y. First, a model must be constructed which contains enough endogenous structure to portray how changes in one variable, say X, affect the present and future values of both X and Y. The model should generate a pattern of oscillatory behavior in Y similar to that observed in real life.[8] Moreover, model behavior should arise primarily from the model's internal structure, not from exogenous inputs driving the model.[9]

Borrowing from an example developed in Section IV, suppose the modeler wants to analyze the impact of delivery delay on sales in a firm which has experienced unstable sales growth. An adequate model to address this question should include the hypothesized direct effect of delivery delay on sales effectiveness and, consequently, sales rate (increasing delivery delay reduces sales effectiveness, while decreasing delivery delay increases sales effectiveness); but the model should also include the influence of sales (that is, orders) on order backlogs and delivery delay, as well as the influence of sales rate on revenues, capacity expansion, and marketing effort. When simulated, the model should generate the pattern of unstable sales growth observed in the firm. Construction and analysis of a model containing the feedback interactions

[8] See Forrester [7]. Chapter 13 discusses particular aspects of oscillatory behavior, such as average periodicity and phase relationships between variables, which provide valid measures of the correspondence between model-generated oscillations and observed oscillations.

[9] See Forrester [7], Chapter 12.

between production capacity, delivery rate, delivery delay, marketing effort and sales rate provides the necessary foundation for analyzing the importance of the hypothesized impact of delivery delay on sales.

The second step in whole-model testing involves simulating the model both with and without the direct influence of X on Y. How does the behavior of the variable Y change, as a result of deleting the direct link between X and Y? In the delivery delay example, assessing the impact of delivery delay on sales behavior would involve omitting the hypothesized direct link between delivery delay and sales, and then seeing whether model behavior is altered significantly as a consequence.

Finally, the third step in whole-model testing involves analyzing the causes of the behavior observed in the second step. If the behavior of Y is relatively unaltered, what other variables appear to dominate the behavior? If Y's behavior is altered significantly, what direct and indirect links between X and Y account for the change in behavior?

Before proceeding to give specific examples of the model-behavior testing process, a short discussion of alternative criteria for model-behavior testing is appropriate. At least three criteria are possible:

(1) Does omission [inclusion] of the factor lead to a change in the predicted numerical values of the system?

(2) Does omission [inclusion] of the factor lead to a change in the behavior mode of the system? (For example, does it damp out or induce fluctuations in the system?)

(3) Does omission [inclusion] of the factor lead to rejection of policies that were formerly found to have had a favorable impact or to reordering of preferences among alternative policies?

In general, the results of an evaluation of the importance of an hypothesized impact of one variable on another will depend on which of the three

criteria above are used. For example, suppose the model designed to explore the causes of sales fluctuations in a particular firm initially exhibits the sales behavior shown by the curve labeled A in Figure 1. Suppose now that omission (or inclusion) of the direct link between delivery delay and sales alters model behavior to that described by Curve B in Figure 1. Curve B differs from A in the exact numerical values for sales over time, but both curves clearly exhibit approximately the same general growth trend and the same magnitude of fluctuations. The difference between outcomes A and B would then be judged important by the first criterion for model-behavior testing given above, but unimportant by the second criterion. To deal with the third criterion, suppose that a number of policies were tested on both the models underlying Curves A and B and it was found that the policies that reduce fluctuations in one model also reduce fluctuations in the other, and conversely. In this situation, the difference between the two models would be

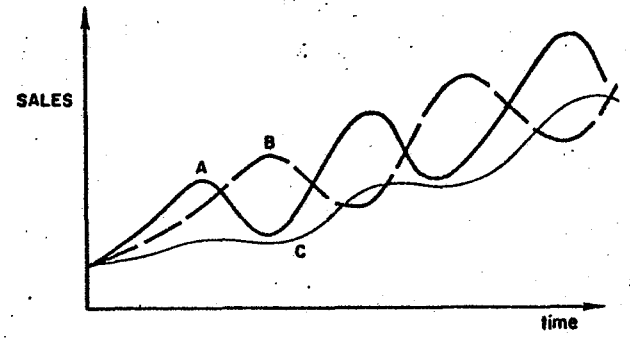

Figure 1. Alternative patterns of sales behavior.

judged insignificant or unimportant by the third criterion.[10] This example, although highly simplified, illustrates some of the considerations involved in assessing the importance of a given model relationship. The purpose of a particular study will, in general, determine which of the three criteria is appropriate for the evaluation process.

IV. A CASE STUDY IN SELECTING MODEL VARIABLES: THE IMPACT OF DELIVERY DELAY ON SALES IN A MODEL OF A FIRM

The following example demonstrates the operation and consequences of the two previously described types of testing. The example shows, first, that model-behavior testing yields information about a variable's influence on behavior that cannot be ascertained by statistical testing; and, second, that the information produced by model-behavior testing can aid in discriminating which variables are important in generating particular patterns of behavior.

---

[10] To show still a more complex case, consider Curve C in Figure 1. Curve C shows alternating periods of growth and leveling off rather than growth and decline as in Curves A and B. Curve C might be judged to be significantly different from, say, Curve A by both the first and second criteria. The outcomes might not be significantly different from the standpoint of the third criterion, however, if , for example, the same policies that reduce fluctuations in A also contract the leveling-off periods in C, and conversely.

A. The Model

The following example utilizes a fairly simple feedback model built by J. W. Forrester [8] to explain how a rapidly growing firm can experience instability in sales even in the presence of potentially limitless demand. Figure 2 provides an overview of the structure of the "market-growth" model. The model assumes that the firm expands or contracts its sales force depending upon

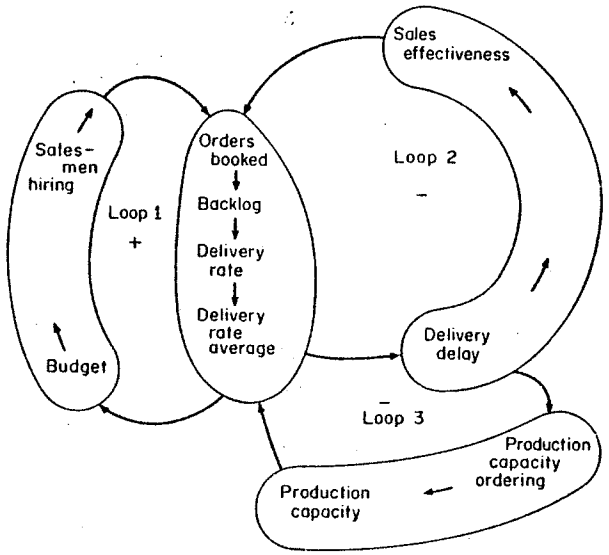

Figure 2. Major feedback loops in Forrester's market-growth model.

the difference between the number of salesmen that can be supported by the marketing budget and the existing sales force. The firm orders additional production capacity when delivery delay becomes longer than desired (the firm's desired delivery delay is taken as two months) and reduces capacity when delivery delay falls below the desired value. Delivery delay also influences sales (orders booked) through sales effectiveness. Assuming a highly competitive market, the model hypothesizes that customers reduce orders whenever delivery delay rises, and vice versa.

Given such a model of interactions within a firm, how could we test the hypothesized impact of delivery delay on sales? Assume that the real-life firm has experienced rapid growth in sales and production capacity punctuated by intermittent sharp declines in sales. Assume also that ample data are available for order backlog, delivery delay as recognized by the market, salesmen, and production capacity.

B. Statistical Tests of the Hypothesized Impact of Delivery Delay on Sales

In order to test statistically the hypothesized impact of delivery delay on sales, one must write out the appropriate equations, identifying unknown parameters to be estimated. The equations given below, taken from the original Forrester model, describe the determinants of change in order backlog, orders booked (sales), and delivery rate. Random terms are included in the equations for orders booked and delivery rate:

$$\Delta BL = OB_t - DR_t \tag{5}$$

$$OB_t = S_t \cdot SE_t + \varepsilon_{1_t} \tag{6}$$

$$SE_t = g_1(DDRM_t) \tag{7}$$

$$DR_t = PC_t \cdot PCF_t + \varepsilon_{2_t} \tag{8}$$

$$PCF_t = g_2(BL/PC)_t \tag{9}$$

$$\Delta BL = BL_{t+1} - BL_t$$

where
- BL = order backlog (units)
- DR = delivery rate (units shipped per month)
- SE = sales effectiveness (units sold per salesman per month)
- PC = production capacity (maximum possible units shipped per month)
- DDRM = delivery delay recognized by market (months)
- PCF = production capacity fraction (dimensionless)
- $\varepsilon_1, \varepsilon_2$ = random error processes
- $g_1(\cdot)$ = nonlinear function
- $g_2(\cdot)$ = nonlinear function

In order to estimate Equation (5), the order backlog equation, the nonlinear functions g_1 (DDRM) and g_2 (BL/PC) must be parameterized. In the following experiments, the function $g_1(\cdot)$, which incorporates the impact of delivery delay on sales, is approximated as linear and $g_2(\cdot)$ is specified as a third-order polynomial ($g_2(\cdot)$ is definitionally constrainted to be zero when (BL/PC) equals zero).[11] After rearranging, delivery delay recognized by market DDRM, the variable whose impact is being tested, enters in the second coterm multiplied by the parameter K2:

[11] Approximating $g_1(\cdot)$ as linear and $g_2(\cdot)$ as a third-order polynomial allows the modeler to draw the maximum statistically significant information from the experimental data. The variation in delivery delay recognized by the market DDRM does not carry that variable into significantly nonlinear regions of the relationship $g_1(\cdot)$. The opposite holds for the backlog to production capacity ratio (BL/PC), which ranges well into the nonlinear regions of $g_2(\cdot)$.

$$\begin{aligned}\Delta BL &= S_t \cdot (K1 + K2 \cdot DDRM_t) + PC_t \cdot (K3 \cdot (BL/PC)_t + K4 \cdot (BL/PC)_t^2 \\ &\quad + K5 \cdot (BL/PC)_t^3) + (\varepsilon_{1t} + \varepsilon_{2t}) \\ &= K1 \cdot S_t + K2 \cdot S_t \cdot DDRM_t + K3 \cdot BL_t + K4 \cdot (BL^2/PC)_t + K5 \cdot (BL^3/PC_t^2) \\ &\quad + \varepsilon_t \end{aligned} \tag{9}$$

$$\Delta BL = BL_{t+1} - BL_t \ .$$

ε_t is assumed to be zero-mean, normally distributed, stationary and white.

K1, ..., K5 are unknown parameters.

According to prior reasoning, increases in delivery delay should suppress sales, hence the sign of K2 should be negative.

To test statistically the hypothesis that delivery delay influences sales, one could examine the t-statistic for the parameter K2 or the partial correlation coefficient for the coterm $S_t \cdot DDRM_t$ in Equation (9). In order to examine the performance of the statistical tests, we conduct the following simple experiment. Simulating the entire market-growth model, we generate synthetic data for the variables involved in Equation (9) and use that data for estimating the equation. Such synthetic data experiments are common in the econometrics and statistics literature on evaluating alternative estimators and have been recently used to evaluate estimators for feedback models (Senge [26]) In the present case, the experimental framework permits examination of the performance of statistical tests of the impact of delivery delay on sales.

First, an experiment is conducted under the highly-idealized conditions of pe fect measurement of all model variables. As shown in the first estimation in Table when data measurement is perfect, statistical estimation (using ordinary least-squa

(OLS)) results in a statistically significant estimate for the parameter K2 (t-statistic equal to -9.69) and a high partial correlation coefficient (-0.7049). However, when moderate measurement errors[12] are permitted to enter the data, the statistical results are adversely affected, as shown in the second estimation in Table 1. The t-statistic (-1.248) indicates that the estimated delivery delay impact

TABLE 1

STATISTICAL TESTS FOR ORDER BACKLOG EQUATION
(Ordinary Least-Squares Estimation)
Error-Free Data

$$BL = BL_{-1} + K1 \cdot S_{-1} + K2 \cdot S_{-1} \cdot DDRM_{-1} + K3 \cdot BL_{-1} + K4 \cdot (BL_{-1}^2/PC_{-1}) + K5 \cdot (BL_{-1}^3/PC_{-1}^2).$$

COEFF	TRUE VALUE	ESTIMATED VALUE	$\hat{\sigma}_{\hat{\beta}}$	T-STAT	r_h	
K1	475	457.7	37.05	12.36	0.7851	
K2	-61.5	-54.62	5.638	-9.69	-0.7049	R^2 =
K3	-0.6178	-0.6484	0.06398	-10.13	-0.7207	0.9934
K4	0.1324	0.136	0.02018	6.74	0.5689	
K5	-0.00975	-0.00975	0.00299	-3.26	-0.3171	

Error-Corrupted Data

COEF	TRUE VALUE	ESTIMATED VALUE	$\hat{\sigma}_{\hat{\beta}}$	T-STAT	r_h	
K1	475	408.2	89.48	4.562	0.4239	
K2	-61.5	-26.85	21.51	-1.248	-0.1270	
K3	-0.6178	-0.5563	0.1530	-3.637	-0.3496	R^2 =
K4	0.1324	0.07719	0.05362	1.440	0.1461	0.8486
K5	-0.00975	-0.00411	0.00606	-0.6792	-0.0695	

[12] Random errors with standard deviations equal to 10% of the current value of the error-free data are present in Table 2. Errors of 5-10% are typical in the economic data according to Morgenstern [19].

is statistically insignificant. Moreover, the coterm $S_t \cdot DDRM_t$ exhibits low partial correlation relative to the partial correlation for the coterms K1, K3, and K4. Overall, the statistical tests based on the error-corrupted data give no evidence to support the hypothesized influence of delivery delay on sales, even though, by the very design of the computer experiment, a direct impact of delivery delay on sales is present in the data-generating model. Therefore, the statistical tests are not reliable tests of the specification of the order backlog equation. What the tests do indicate is that, when the quality of available data becomes poor, the hypothesized impact of delivery delay on sales becomes difficult to estimate. In this sense, the statistical tests measure the usefulness of the available data rather than the specification of the market-growth model.

C. Model-Behavior Test of the Hypothesized Impact of Delivery Delay on Sales

The preceding section showed that, even with perfect knowledge of the determinants of sales, statistically insignificant estimates of the influence of delivery delay on sales may be obtained when data contain moderate measurement errors. An econometrician viewing this result might conclude that delivery delay is a relatively unimportant influence on sales. However, this section shows, using model-behavior testing, that delivery delay actually exerts a pronounced influence on sales behavior in the market-growth model.

Figure 3 shows the basic behavior of the market-growth model when the direct link between delivery delay and sales, shown in Figure 2, is present.

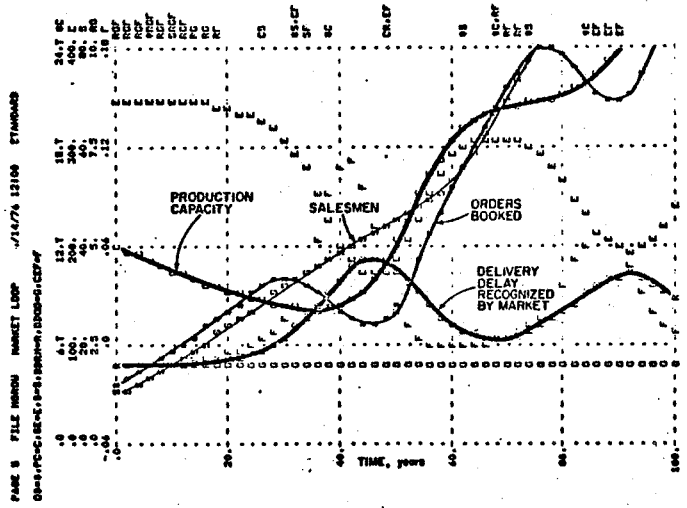

Figure 3. Standard run of market-growth model.

(To simplify the discussion, the simulations in this section do not include random components.) Figure 3 shows a behavior pattern of growth in sales (orders booked), interrupted by periods of decline. Such growth instability is caused by the simultaneous effect of delivery delay on sales and capacity expansion. At the outset of the simulation, the firm's delivery delay is low. Sales effectiveness is consequently high, thereby allowing orders booked to increase rapidly as the sales force increases. The growth of sales increases the firm's order backlog. Delivery delay rises during this period because order backlog is growing more rapidly than production capacity. Eventually, delivery delay increases to the point where high delivery delay begins to suppress orders. As the firm perceives the long delivery delay, it

begins to order new production capacity. However, due to delays in perceiving delivery delay and in acquiring production capacity, delivery delay as perceived by the market continues to rise for a while, thereby lowering sales effectiveness and depressing orders booked. As capacity eventually begins to expand, delivery delay starts to decline. The decline in orders booked and order backlogs further lowers delivery delay. Once delivery delay falls, growth in orders booked resumes and the rate of capacity expansion declines. Figure 3 therefore shows alternating periods of growth and decline in sales and production capacity. In the following discussion, the pattern of system behavior shown in Figure 3 will be assumed to be the pattern of real-life behavior which the modeler seeks to reproduce and understand.

In order to test the contribution of delivery delay to sales behavior, the direct impact of delivery delay on salesman effectiveness can be eliminated from the market-growth model. A simulation of the resulting model is shown in Figure 4.[13] In contrast to the previous simulation, Figure 4 shows continued growth in orders booked and continued increase in production capacity after month 30, although both orders and capacity expand at a fluctuating rate of growth. Therefore, with the direct link between delivery delay and sales eliminated, the market-growth model no longer exhibits the recurring fall-offs in sales characteristic of the original model. With this link omitted, a high delivery delay no longer exerts a depressive effect on sales. Instead, sales are influenced only by the level of salesmen and by a constant sales effectiveness.

---

[13] Note that some vertical plot scales are different in Figures 3 and 4 due to the expanded range of variation in salesmen, orders booked, and production capacity in Figure 4.

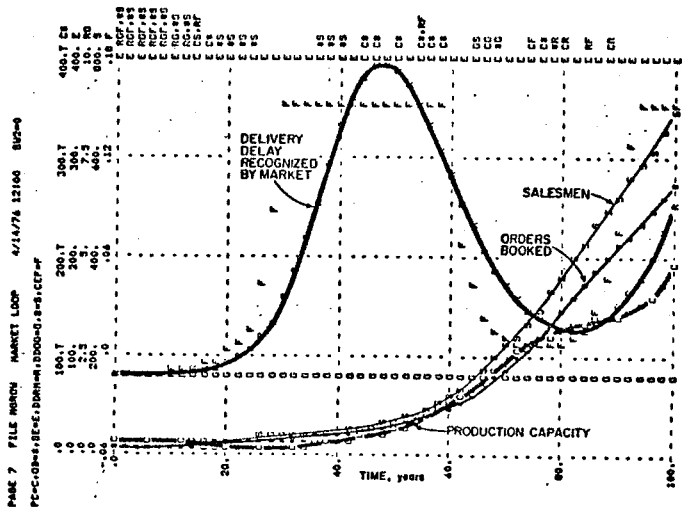

Figure 4. Market-growth model with no direct influence of delivery delay on orders booked (sales).

Analysis of the basic feedback structure of the market-growth model, depicted in Figure 2, clearly shows why omission of the influence of delivery delay on sales shifts model behavior from the pattern seen in Figure 3 to the outcome in Figure 4. Loop 1 in Figure 2 is a positive feedback loop. Within the loop, an increase in delivery rate raises revenues; high revenues, in turn, will increase the marketing budget, leading to increased salesmen, increased orders, and, thereby, still higher sales. Loop 1, therefore, tends to promote continued exponential growth in sales.

Loops 2 and 3 are negative feedback loops controlling delivery delay. In Loop 2, an increase in orders raises order backlog, causing delivery delay

to increase. In turn, a higher delivery delay lowers salesman effectiveness, thereby lowering sales.

In loop 3, an increase in delivery delay encourages capacity expansion, causing an increase in delivery rate and a decline in delivery delay.

When delivery delay is omitted as a direct influence on sales rate, Loop 2 no longer constrains sales growth. Sales behavior, instead, is controlled by the positive feedback loop, Loop 1. Inactivation of Loop 2, therefore, accounts for the nearly steady exponential growth in sales observed in Figure 4. In Figure 4, production capacity grows continually after month 30 (due to steadily rising sales), but at a fluctuating rate of increase. These fluctuations in production-capacity expansion are generated in Loop 3, and are coupled to the fluctuations in delivery delay seen in Figure 4.

Comparison of Figures 3 and 4 shows the critical impact of delivery delay on sales behavior. When delivery delay is no longer assumed to influence sales, the behavior of the market-growth model is altered markedly, and the model no longer exhibits alternating priods of growth and decline in sales. In other words, the overall behavior mode of the system is significantly changed when the delivery delay effect is omitted. Model-behavior testing, therefore, shows that the hypothesized impact of delivery delay on sales exerts an important influence on model behavior. This result is especially striking in light of the outcome of the statistical tests in Table 1, which showed that delivery delay was not a statistically significant determinant of sales. The outcomes of the alternative tests highlight the point that only

the model-behavior test can be viewed as a true measure of the importance of the hypothesized impact of delivery delay on sales.[14]

V. APPLICATIONS OF MODEL-BEHAVIOR TESTING TO TEST ALTERNATIVE THEORIES OF SOCIAL BEHAVIOR

The two previous sections have outlined the principal components of the model-behavior testing approach and given a detailed example of its application. This section of the paper summarizes an application of whole-model testing in analyzing an economic system, and outlines a number of areas where whole-model testing may help to clarify the merits of alternative theories of social behavior.

[14] An important issue which cannot be explored in so short an example is the sensitivity of the model-behavior test outcome to variations in model parameters. One obvious parameter to consider is the production-capacity receiving delay--the length of time required to obtain new production capacity. If the delay is shortened (the delay initially equals 12 months), capacity can be acquired (or reduced) more quickly, thus leading to more rapid capacity expansion and sales growth in the full model (with the delivery-delay impact on order rate). However, unless the delay time is less than one month (an unrealistically low value even for labor acquisition, given typical hiring and training delays), the model still generates a pattern of oscillating sales growth. If the production capacity receiving delay is lengthened, the model's oscillatory behavior will become even more severe until, in the limit of infinite capacity receiving delay (constant production capacity), sales oscillate about an equilibrium. So long as the model still oscillates, omission of the delivery delay - order rate link will have a pronounced effect on model behavior. Therefore, the model-behavior testing outcome is quite insensitive to variations in the length of the capacity acquisition delay. In applying model-behavior testing, considerable attention should be devoted to such sensitivity testing (see Mass [17] for example).

In a recent application, Mass [17, 18] used whole-model testing to assess the generic causes of short-term and medium-term cycles in a national economy. In the study, he develops a general model of a producing unit within the economy, and then modifies the production model to incorporate various hypothesized causes of economic instability.

One particularly significant outcome of the study concerns the relative importance of labor adjustments and fixed capital investment in generating short-term business cycles. As noted by Burns [4], the predominant number of business-cycle theories, including the theories of Paul Samuelson, John Hicks, Nicholas Kaldor, and James Duesenberry, emphasize fluctuations in fixed capital investment as a cause of overall fluctuations in income and output.[15] Such theories have been widely influential from a theoretical standpoint, and have stimulated much subsequent business-cycle research. They have gained further support in the form of statistical studies showing relatively large swings in investment over a typical business cycle (see, for example, Evans [6]). Moreover, widespread acceptance of the theories has led to economic stabilization policies which emphasize regulating investment opportunities.

Mass employs model-behavior testing to evaluate the widely-held hypothesis that fluctuations in capital investment are essential to the business cycle. In his first test, Mass holds fixed capital stock constant while labor is allowed to vary. The resulting simulation exhibits a four-year fluctuation resembling the short-term business cycle in terms of amplitude, phase relationships between variables, and other characteristics. This simulation indicates

[15] See Samuelson [23], Hicks [13], Kaldor [14], and Duesenberry [5] for the original statement of these theories.

that a short-term business cycle can be generated independent of fluctuations in fixed capital investment.

In a second test, labor (employment) is held constant while capital stock is permitted to vary. The resulting simulation exhibits a fluctuation of around eighteen-year periodicity resembling the so-called Kuznets cycle (see Abramovitz [1]). The outcome suggests that variations in fixed capital investment alone, without variations in employment, cannot account for the occurrence of short-term business cycles in the economy, and probably underlie much longer-term cycles.

Similar model-behavior testing might be employed to address some of the major controversies in economic theory and policy. For example, most attempts to date to evaluate the monetarist theories regarding economic cycles and stabilization policy have utilized single-equation statistical tests (for example, Andersen and Jordan [2]). Reflecting considerations similar to those described above, Blinder and Solow [3] have argued that the single-equation approach has the theoretical weakness that it ignores the feedback from changes in income to changes in fiscal and monetary policy. A more fruitful approach for evaluating the monetarist theories might be to incorporate the monetarist assumptions into a broad national model to evaluate their implications, significance, and interaction. Such a model might draw, for example, on Friedman's 1968 Presidential Address to the American Economics Association (Friedman [11]), in which he presents a descriptive summary of some of the main relationships linking money supply, interest rates, GNP, prices and price expectations, and capital investment according to the monetarist view. Such a model should help to define and assess these assumptions in a more comprehensive framework than has been available to date.

Numerous other potential applications of model-behavior testing reside in psychology, sociology, education, and the other social sciences. The model-behavior approach might be used, for example, to examine the effects of motivation, expectations, job availability, and social conditions on the effectiveness of education in a school system or community.[16] Such variables are connected through complex feedback loops of cause-and-effect relationships, and it is doubtful that their relative importance can be assessed by conventional statistical testing. In complex problem areas, such as economic theory or educational effectiveness, the model-behavior approach affords an unprecedented opportunity for testing the effects of various assumptions and theories of social behavior.

VI. CONCLUSIONS

This paper has examined two approaches for determining whether or not an hypothesized impact of one variable upon another should be included in a model. The first method analyzed was the single-equation statistical testing approach. The second approach entailed the analysis of model behavior. The major finding of the paper is that, of the two basic approaches, only behavior tests provide a valid basis for selecting model variables. Only by analysis of model behavior can the modeler ascertain the importance of a particular variable. He can do so by omitting any influence of the variable from the model, or by constraining the variable's movement, and examining the consequent shift

[16] Foster [10] and Roberts [22] have conducted preliminary efforts along these lines.

in model behavior. Model-behavior testing can be used to isolate the influence of an individual variable on a particular historical behavior pattern, on a possible mode of future behavior, or on model response to alternative policies.

By contrast, the proper role of single-equation statistical tests is much narrower. For example, the widely-used t-test and partial correlation coefficient provide information only on the precision with which a given parameter can be estimated, not on the importance of the parameter or the associated variables. An example in Section IV showed how an hypothesized impact of delivery delay on sales in a firm could fail the statistical tests, even though the delivery delay impact was crucial for the behavior of the particular model in question. Given the narrower focus of the statistical tests, they should be viewed more as tests of data usefulness than as tests of model specification. That is, failure to pass the statistical tests should not lead the modeler to reject the hypothesized relationship in question, but rather to recognize that the hypothesis is difficult to measure from the available data. Conversely, passing such tests does not mean that the hypothesis in question is in any sense "important," only that it is measurable. If indicators of measurability are used as guides to model specification, they can lead to rejection of relationships that are extremely important for system behavior.

Future research should endeavor to delineate further the possible uses and misuses of statistical and model-behavior testing. A number of theoretical debates in the social sciences might be resolved or at least greatly clarified through application of the behavior testing approach outlined here.

The National Model currently being constructed in the System Dynamics Group at MIT (see Forrester, Mass, and Ryan [9]) should provide a powerful framework for analyzing many of the more persistent controversies in economics, such as the role of capital investment and monetary policy in economic stabilization, and the monetarist-fiscalist debates. Future research should also focus on possibilities for integrating the two testing approaches. Are there, for example, circumstances under which the results of single-equation tests provide a useful input to behavior testing? Such questions can probably only be answered within the context of a well-defined validation problem, such as the comparison of alternative investment formulations currently being conducted as part of the above-mentioned National Modeling Project at MIT (see Senge [27]).

Continued research aimed at developing the model-behavior testing approach and integrating it more fully with more established testing approaches may contribute to a basic reorientation of model building and theory testing in the social sciences. Although the limitation of statistical testing have been well understood for some time, modeling practice continues to be dominated by the statistical testing perspective. Model-builders continue to reject hypotheses on the basis of low statistical significance (see Goldfeld [12], Schultz [24], for example), and to attribute inconclusive statistical studies to poor data bases, as if more complete data would enable the statistical testing methodology to discriminate successfully among alternative hypotheses. This continuing reliance on statistical tests is closely linked to the heritage of single-equation models common to all the social sciences. To the extent that researchers still construct single-equation

models, model-behavior testing is not possible. In fields such as econometric modeling, the emergence of system models has preceded the adoption of a commensurate approach to model testing; consequently, econometricians tend to build multiple-equation models within an essentially single-equation philosophy of model testing. This is unfortunate, for such practice overlooks one of the greatests strengths inherent in the systems approach--to test the effect of alternative hypotheses on a complete system, just as is done in the experimental sciences. Hopefully, increased understanding of basic issues such as those raised in this paper will begin to foster a new philosophy and approach for theory testing in the social sciences.

BIBLIOGRAPHY

[1] Abramovitz, Moses, "The Nature and Significance of Kuznets Cycles," Economic Development and Cultural Change, vol. 9 (April 1961).

[2] Andersen, L. C., and J. L. Jordan, "Monetary and Fiscal Actions: A Test of their Relative Importance in Economic Stabilization," Federal Reserve Bank of St. Louis Review, vol. 52 (November 1968), pp. 11-24.

[3] Blinder, A. S., and R. M. Solow, "Analytical Foundations of Fiscal Policy," in The Economics of Public Finance (Washington: The Brookings Institute, 1974), pp. 67-70.

[4] Burns, Arthur F., The Business Cycle in a Changing World (New York: National Bureau of Economic Research, 1969).

[5] Deusenberry, James S., Business Cycles and Economic Growth (New York: McGraw-Hill, 1969).

[6] Evans, Michael K., Macroeconomic Activity (New York: Harper and Row, 1969).

[7] Forrester, Jay W., Industrial Dynamics (Cambridge: MIT Press, 1961).

[8] Forrester, Jay W., "Market Growth as Influenced by Capital Investment," Industrial Management Review, vol. 9, no. 2 (Winter 1968).

[9] Forrester, Jay W., Nathaniel J. Mass, and Charles J. Ryan, "The System Dynamics National Model: Understanding Socio-Economic Behavior and Policy Alternatives," Technological Forecasting and Social Change, vol. 9 (July 1976).

[10] Foster, Richard O., "Education in the City," System Dynamics Group Working Paper D-2144, Alfred P. Sloan School of Management, MIT (Cambridge: 1972).

[11] Friedman, Milton, "The Role of Monetary Policy," American Economic Review, vol. 58 (January 1968), pp. 1-17.

[12] Goldfeld, Stephen M., "The Demand for Money Revisited," Brookings Papers on Economic Activity, 3:1973.

[13] Hicks, John R., "Mr. Harrod's Dynamic Theory," Economica, vol. 16 (May 1949), pp. 10-121.

[14] Kaldor, Nicholas, "A Model of the Trade Cycle," Economic Journal, vol. 50 (March 1940), pp. 87-92.

[15] Kalman, Rudolf E., "A New Approach to Linear Filtering and Prediction Problems," Journal of Basic Engineering, Series D, vol. 82 (March 1960).

[16] Keynes, J. M., "Professor Tingerben's Method," [a review of J. Tinbergen, Statistical Testing of Business Cycle Theories: A Method and Its Application to Investment Activity (Geneva: League of Nations, 1939)], Economic Journal (September 1939), pp. 567, 569.

[17] Mass, Nathaniel J., Economic Cycles: An Analysis of Underlying Causes (Cambridge: Wright-Allen Press, 1975).

[18] Mass, Nathaniel J., "Modeling Cycles in the National Economy," Technology Review, vol. 78 (March-April 1976).

[19] Morgenstern, O., On the Accuracy of Economic Observations (Princeton: Princeton University Press, 1963).

[20] Morrison, Denton E. and Ramon E. Henkel, eds., The Significance Test Controversy (Chicago: Aldine Publishing Co., 1970).

[21] Peterson, David W., Hypothesis, Estimation, and Validation of Dynamic Social Models (Ph.D. Thesis, Department of Electrical Engineering, MIT, Cambridge, Mass., 1975).

[22] Roberts, Nancy, "A Computer System Simulation of Student Performance in the Elementary Classroom," Simulation and Games, vol. 5 (September 1974), pp. 265-289.

[23] Samuelson, Paul A., "Interactions Between the Multiplier Analysis and the Principle of Acceleration," Review of Economics and Statistics, vol. 21 (May 1939), pp. 75-78.

[24] Schultz, T. Paul, "An Economic Model of Family Planning and Fertility," Journal of Political Economy, vol. 77, no. 2 (April 1969).

[25] Schweppe, Fred, Uncertain Dynamic Systems (Englewood Cliffs, N.J.: Prentice Hall, 1973).

[26] Senge, Peter M., "Testing Estimation Techniques for Social Models," System Dynamics Group Working Paper D-2199-4, Alfred P. Sloan School of Management, MIT (Cambridge: 1975).

[27] Senge, Peter M., "The System Dynamics National Model Investment Formulation: A Comparison to the Neoclassical Model," System Dynamics Group Working Paper D-2431, Alfred P. Sloan School of Management, MIT (Cambridge: 1976).

[28] Theil, H., Principles of Econometrics (New York: John Wiley and Sons, 1971).

[29] Zellner, Arnold, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," Journal of the American Statistical Association, vol. 57, pp. 977-992 (1963).