RESEARCH PROBLEMS:

In this department DYNAMICA presents problems that have the potential to stimulate research involving the system dynamics perspective. Articles may address real-world dynamic problems that could be approached fruitfully using system dynamics, or methodological problems affecting the field. A submitted paper should concisely motivate and define a problem and start a process of conceptualization or formulation that can open the way for further studies. Manuscripts, not exceeding 2,000 words, should be sent to George P. Richardson, System Dynamics Group, E40-294, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, U.S.A.

USING FEEDBACK MODELS TO TEST THE ROBUSTNESS OF STATISTICAL PROGRAM EVALUATION DESIGNS

Problem submitted by David F. Andersen
Graduate School of Public Affairs, Nelson A. Rockefeller College of Public Affairs and Policy, State University of New York.

Recent experiences with the statistical evaluation of social programs suggest that when the programs being evaluated have impacts that are dynamic and possibly involve feedback effects, the statistical models being used to evaluate the programs may produce conclusions that are not fully justified. This paper proposes a research program that will use feedback simulation models to test the robustness of proposed evaluation designs *before* data collection begins.

THE PROBLEM IN BRIEF

The recent literature evaluating large scale, comprehensive, and long term attempts at social reform is filled with suggestions that such attempted reforms are not effective. Although several compelling examples illustrate the alleged ineffectiveness of social reforms, this paper will concentrate on the example of the alleged ineffectiveness of public schools' attempts to improve student achievement. As noted in the references, this example probably points to a larger problem that plagues most all statistical evaluations of social programs. For example, in a recent survey of over fifteen years of research covering scores of evaluations, Hanushek concluded that attempts to improve student achievement in US schools via public expenditures for educational inputs such as better teachers and better schools have not been effective.[1] These pessimistic results, couched in terms of statistically insignificant or null results, also appear to be accumulating with respect to other areas such as OSHA's regulation of the work force.[2, 3] as well as evaluation of health care.[4]

The amount of time and effort spent collecting data and statistically evaluating these programs is immense. For example, the 130 experiments surveyed by Hanushek are to a large degree replications of and extensions to a basic study of Coleman and others[5] which at the time of its inception was one of the largest and most comprehensive evaluations of schooling ever undertaken. However, several threads of evidence at least suggest that these large scale evaluations (as well as numerous small-scale evaluations) may all be methodologically flawed and that the widely reported null results may in part be mathematical artifacts rather than real policy conclusions. Specifically, if the real world policy system being evaluated is a dynamic system endogenously driven by feedback, then the statistical tools being used to evaluate such programs may be biased toward producing null results.

If in fact the program designs being used to evaluate social programs are biased toward producing null results under certain circumstances, then this implicit methodological bias needs to be more fully investigated and specific techniques proposed for evaluating when and under what conditions such biases may exist. For example, preliminary experiments discussed below suggest that measurement error in independent variables, the existence of non-linearities, or sampling frequency may dramatically bias the quality of estimates of parameters within a dynamic feedback driven system.

In addition to continuing research into the general limits of statistical designs for evaluating social programs, additional research is needed into how feedback simulation models of specific social programs can be used to evaluate a proposed statistical design even before data have been collected. That is, simulation models might be used to generate synthetic data sets to be analyzed by a statistical evaluation model *before* such evaluation models are used to examine real data sets. If the evaluation model cannot correctly analyze the synthetic data generating system, the analyst should doubt the ability of the evaluation model to correctly analyze a real system.

BACKGROUND TO THE PROBLEM

There would appear to be at least four relevant bodies of literature that could be drawn into this research problem two from the field of statistical evaluation research and two from the field of dynamic modeling.

First, since this whole problem is motivated by the discovery of null results, one should probably become familiar with some of these substantive research findings. Hanushek's

review of the schooling effectiveness literature would be a good place to start as might be the OSHA and health care cases cited above.

A second body of literature would explore the "classic" methods of statistical evaluation design beginning with the basic linear regression model[6] and several of its extensions in econometrics.[7] In addition, considerable discussion exists concerning the limits of and difficulties in applying such models to specific substantive areas. For example, in the field of schooling effectiveness research, one might look at the work of Hanushek,[8] Bridge, Judd, and Hoock,[9] or Bidwell and Windham.[10] Finally, general techniques for exploring the sensitivity of regression results to violations of basic assumptions such as Belsley, Kuh, and Welch's interesting work[11] would also be explored in this body of research.

A third review would look into the statistical properties of feedback driven systems with special emphasis on how data extracted from such systems might conform to or violate the assumptions implicit in the basic linear regression model and its most popular extensions. Feedback driven systems, such as the simulation models constructed in most system dynamics models, have statistical properties that are somewhat different from the properties assumed by ordinary least squares regression. Here some of Kalman's original papers might be useful[12, 13] as well as the exposition in any standard text in the field of stochastic estimation of dynamic systems.[14] Several papers also exist that attempt to compare and contrast filtering techniques within feedback driven models with more classic regression based analyses.[15, 16]

Finally, and perhaps most important, a body of literature on synthetic data experiments empirically explores some of the limits of classical regression analysis when applied to dynamic systems. In these experiments, a dynamic model with stochastic inputs has been used to generate a synthetic data set. A regression based estimation model is then used to attempt to recover the parameters known to exist within the synthetic data generating system. Since the structure and parameters of the data generating "reality" are fully known, the performance of the estimating model can be accurately assessed.

For example, Senge has used this experimental design to investigate how well regression models can recover parameters from a dynamic system when the observed independent variables have been corrupted by relatively small amounts of measurement error.[17] For feedback driven systems, the estimation models performed very well for no or minute quantities of measurement error, but the estimation model's performance degenerated rapidly under small to moderate amounts of measurement error. Along a similar vein, Mass and Senge explored the ability of statistical tests of significance to recover important controlling parameters within non-linear data generating models.[18] Standard statistical tests showed that parameters known to be critical determinants of overall model performance were not significant.

In a most interesting experiment, Luecke and McGinn[19] used a modified Markov chain to investigate exactly the schooling effectiveness question surveyed by Hanushek. Luecke and McGinn discovered that even when expenditures for educational inputs such as better teachers and schools are known to have an effect on student achievement in a simulated reality, estimation models similar to those reviewed by Hanushek still produced null results. More than any other of the synthetic data experiments, this result questions whether the null results from statistical evaluations are real policy conclusions or mathematical artifacts induced by the evaluation design being employed. Andersen has more fully discussed these points elsewhere.[20]

POSSIBLE APPROACHES TO THE PROBLEM

One possible approach to studying the robustness of program evaluation designs would follow a line of research begun by Richardson.[21] While investigating the bias introduced by discrete sampling of data from a continuous system, Richardson mathematically derived an estimate of the bias in parameter estimates that would result from various sampling strategies. Similar mathematically based derivations of the stochastic properties of feedback driven systems could be undertaken.

A second, more empirically oriented approach would follow on the synthetic data experiments reported above and attempt to isolate under what types of circumstances the statistical estimation of feedback systems breaks down. Such an empirical approach would have less of the flavor of a proof than the Richardson approach, but could still provide generally applicable guidelines for knowing when certain evaluation designs *might* prove to be flawed.

A final approach would be to propose a procedure for testing the robustness of a specific evaluation design for a specific system. For example, a researcher interested in investigating a specific hypothesis using a specific evaluation design would construct two or more synthetic models of the system under investigation before any data collection takes place. Within one (or several) of these synthetic models the proposed hypothesis would hold and within the other one (or several) the proposed hypothesis would not be obtained.

For example, a researcher might construct two simulation models of the effects of government inspections on injury rates in industrial workplaces. One model would *assume* that government inspections decrease injury rates. The second model would *assume* that government inspections have no effect. The proposed evaluation design would then be used to analyze data generated from these two synthetic realities. If the proposed evaluation design showed significant statistical results when the hypothesis was known to exist and failed to show significant results when such hypothesized effects did not in fact exist in the synthetic reality, then the researcher could have greater confidence in the robustness of the research design *before* real data collection began. In the example just cited if the simulation model assumed that the government inspections mattered and an analysis of this simulation's output confirmed that such inspections mattered, then confidence in the statistical research design would increase. Similarly, if the simulation model assumed that government inspections made no difference and statistical analysis of this simulation's output found no significant effect, then confidence in the statistical research design would be further increased. In those cases where statistical analysis failed to

"discover" assumptions built into the simulation model, confidence in the statistical research design would decrease.

The key to such a research program would be in designing a set of generic procedures or tests which, if passed, would satisfy most observers that the proposed evaluation design is relatively robust and not subject to flaws that could easily be detected before data gathering. Such a research program could also suggest how iterative use of such synthetic data experiments could be used to improve the evaluation models as well as test sensitivity to possible types of specification error in the overall evaluation design. Of course, such experiments could not prove that an evaluation design is truly correct or unbiased. However, these experiments could demonstrate that an evaluation design is not subject to a range of possible flaws that are entirely diagnosable *before* data collection begins.

SIGNIFICANCE OF THE STUDY

The significance of this study stems from the very real possibility that policy decisions to fund or defund programs may be based on evaluations whose results are mathematical artifacts rather than statements about the real world. The ability to test the theoretical robustness of proposed evaluation designs before data collection begins could help policy analysts to disentangle better real policy conclusions from possible mathematical artifacts.

REFERENCES

1. HANUSHEK, E.A., (1981), "Throwing Money at Schools," *Journal of Policy Analysis and Management,* 1 (1).
2. SMITH, R., (1979), "The Impact of OSHA Inspections on Manufacturing Injury Rates," *Journal of Human Resources,* 14.
3. McCAFFREY, D., Forthcoming (1983), "An Assessment of OSHA's Recent Effects on Injury Rates," *Journal of Human Resources.*
4. MECHANIC, D., (1979), "Correlates of Physician Utilization: Why Do Major Multivariate Studies of Physician Utilization Find Trivial Psychosocial and Organizational Effects?" *Journal of Health and Social Behaviour.*
5. COLEMAN, J.S., et. al., (1966), *Equality of Educational Opportunity,* U.S. Department of Health, Education and Welfare: Office of Education, Washington, D.C.
6. NETER, J., and W. WASSERMAN, (1974), *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs,* Richard D. Irwin, Inc., Homewood, Illinois.
7. WONNACOTT, R.J., and T.H. WONNACOTT, (1970), *Econometrics,* J. Wiley, New York.
8. HANUSHEK, E.A., "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources,* 14:3.
9. BRIDGE, R.G., M. JUDD, and P.R. MOOCK, (1979), *The Determinants of Educational Outcomes,* Ballinger, Cambridge, Ma.
10. BIDWELL, C.E., and D.M. WINDHAM, (1980), *The Analysis of Educational Productivity, Volume II: Issues in Micro-analysis,* Ballinger, Cambridge, Ma. – see especially D. Rogesa, "Time and Time Again: Some Analysis in Longitudinal Research".
11. BELSLEY, D.A., E. KUH, and R.E. WELCH, (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity,* Wiley, New York.
12. KALMAN, R.E., (1960), "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Engr. (Trans. ASMF),* 820: 35-42.
13. KALMAN, R.E., (1978), "A Retrospective after twenty years: From the Pure to the Applied," in *Applications Of Kalman Filters to Hydrology Hydraulics, and Water Resources* (C.L. Dhiu, ed.), Stochastic Hydraulics Program, Department of Civil Engineering, University of Pittsburgh.
14. GELB, A. (ed.), (1974), *Applied Optimal Estimation, M.I.T. Press, Cambridge, Ma.*
15. PETERSON, D.W., (1980), "Statistical Tools for System Dynamics" in *Elements of the System Dynamics Method* (Jorgen Randers, ed.), M.I.T. Press, Cambridge, Ma.
16. ANDERSEN, D.F. (1981), "Kalman Filter Estimation of System States Compared to the General Regression Problem," *International Journal of Policy Analysis and Information Systems,* 5 (2): 95-109.
17. SENGE, P.M., (1977), "Statistical Estimation of Feedback Models," *Simulation,* 28 (6).
18. MASS, N.J., and P.M. SENGE, (1980), "Alternative Tests for Selecting Model Variables," in *Elements of the System Dynamics Method* (Jorgen Randers, ed.), M.I.T. Press, Cambridge, Ma.
19. LUECKE, D. and N. McGINN, (1975), "Regression Analyses and Education Production Functions: Can They be Trusted? *Harvard Educational Review, 44.*
20. ANDERSEN, D.F., (1980), "How Differences in Analytic Pardigms Can Lead to Differences in Policy Conclusions – Two Case Studies" in *Elements of the System Dynamics Method* (Jorgen Randers, ed.), M.I.T. Press, Cambridge, Ma.
21. RICHARDSON, G.P., (1981), "Statistical Estimation of Parameters in a Predator-Prey Model: An Exploration Using Synthetic Data," July 1981, *System Dynamics Group Memo D-3314-1,* System Dynamics Group, Sloan School of Management, M.I.T., Cambridge, Ma. 02139.