Ex Ante Policy Evaluation, Structural Estimation, and Model Selection

By KENNETH I. WOLPIN*

A number of major social policy interventions have been introduced recently in the United States. The new Temporary Assistance for Needy Families (TANF) program, introduced in 1996, was advertised as changing the welfare system "as we know it." The new Medicare prescription drug benefit, introduced in 2006, was the largest expansion of Medicare in its history. Developing countries are also fertile ground for innovative new policies. Mexico introduced a program in 1997 (Progresa) that provided large subsidies to poor rural households contingent upon the school attendance of their children.

The distinction between ex post and ex ante policy evaluation is important. Ex post policy evaluation occurs upon or after the policy has been implemented. It is ubiquitous in the social sciences. Such studies make use of existing policy variation. Examples include the study of minimum wage effects on labor market outcomes, the study of the impact of welfare benefits on labor market and demographic outcomes, and the study of how divorce laws affect marital stability. The development of methodological approaches to ex post program evaluation using nonexperimental methods is an active area of research (Petra Todd 2006).

There is little methodological or applied research explicitly concerned with ex ante policy evaluation using nonexperimental methods, which is perhaps surprising given its potential value. Interventions that require ex ante evaluation are those that are outside the historical experience. These include a "large" change in the parameters of existing programs such as doubling the (real) minimum wage, adding new features to an existing program such as the Medicare drug benefit program, or introducing a completely new program such as Progresa. The nonexperimental approach to ex ante policy evaluation must be an extrapolation from existing policy or policy-relevant variation.[1] Because of that, and unlike ex post evaluation, ex ante evaluation must rely on parametric and/or behavioral assumptions (theory).

I. An Illustration

Consider performing an ex ante evaluation of the Mexican school attendance subsidy program. I distinguish two hypothetical cases: (a) school tuition, the direct price of schooling, p, varies exogenously across geographic areas from a value between $\underline{p}$ (> 0) and $\bar{p}$; (b) schools are free, $p = 0$. In the first case, it is possible to estimate a relationship between school attendance and tuition cost. In the second case, it is not.

The value of having tuition variation is that a subsidy is simply negative tuition. Suppose that the researcher has decided that in addition to p, school attendance, $s \in \{0, 1\}$ also depends on a set of observed factors, X, and suppose that the data are rich enough to estimate the relationship $s = f(p; X) + \varepsilon$, nonparametrically. It is possible, then, to estimate the effect of the subsidy (τ) on s for all households i in which tuition net of the subsidy, $p_i - \tau$, is in the support of p. Because some values of net tuition must lie outside of the support, it is not possible to estimate the entire response function or to obtain population estimates of the impact of the subsidy. That would require making a parametric assumption about the way tuition enters the response function.

†*Discussant:* Chris Sims, Princeton University

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104 (e-mail: wolpink@ssc.upenn.edu). The author is grateful for support from National Science Foundation grant SES-0450418. I have benefitted from a number of discussions with Aureo DePaula, Donghoon Lee, Kyungchul Song, Frank Schorfheide, and Petra Todd. My thinking on these issues has been significantly influenced by my collaborations with Michael Keane.

[1] The discussion of ex ante policy evaluation relies heavily on Jacob Marschak (1953).

Absent tuition variation, more than just a parametric assumption on f is required. The issue I would like to focus on is how, in practice, a researcher would proceed to develop and "test" a model that can generate an ex ante policy prediction. To illustrate, consider a sequential three-period model.[2] In the model, in each of the first two periods, parents are assumed to choose whether to have their (only) child attend school ($s_t \in \{0, 1\}$ for $t = 1, 2$) or work at a per-period wage, w. In each of those periods, household utility depends on consumption and on the child's school attendance. The utility the parents receive from the child's school attendance is stochastic. The child leaves home after period two. In period three, parents obtain utility from the number of periods the child has attended school (completed schooling), $S_3 \in \{0, 1, 2\}$, and on household consumption, C_t. Utility is linear in consumption and additively separable, namely

$$(1) \qquad U_t = C_t + \varepsilon_t s_t \qquad \text{for } t = 1, 2;$$

$$U_t = C_t + \alpha S_t \qquad \text{for } t = 3,$$

where ε is a serially independent, mean zero normal variate with standard deviation σ. Consumption in period one and two is equal to the sum of parents' income and the child's wage (if the child works), and is equal to parental income in period three, namely

$$(2) \qquad C_t = Y + w(1 - s_t) \qquad \text{for } t = 1, 2;$$

$$C_t = Y \qquad \text{for } t = 3.$$

Parental income and the child wage are constant over time, but vary cross-sectionally. In addition, the child wage is assumed to be the same for all children living in the same village, but to vary among villages. The parents choose s_t in periods one and two to maximize the expected present discounted value of remaining lifetime utility, $V_t = \max(V_t^0, V_t^1)$ for $t = 1, 2$, where V_t^0 is the expected present value if $s_t = 0$, and V_t^1 is the expected present value if $s_t = 1$. Solving backward, the decision rule in period two is

$$(3) \qquad s_2 = 1 \quad \text{iff} \quad \frac{\varepsilon_2}{\sigma} \geq \frac{w - \delta\alpha}{\sigma}; s_2 = 0$$

otherwise,

where δ is the discount factor. The likelihood function is a standard probit. Cross-sectional information in period two on school attendance and on the village-level child wage is sufficient to identify σ and $\delta\alpha$.[3] One cannot separately identify δ and α. It is also not necessary in order to predict the effect of an attendance subsidy on the attendance rate in period two.

Given an attendance subsidy of τ, the household budget constraint becomes

$$(4) \qquad C_t = Y + w(1 - s_t) = \tau s_t \qquad \text{for } t = 1, 2$$

and the term after the inequality in the decision rule (3) is now $(w - \delta\alpha - \tau)/\sigma$. Thus, the effect of the subsidy on the attendance rate is $\Phi[(w - \delta\alpha)/\sigma] - \Phi[(w - \delta\alpha - \tau)/\sigma]$, where Φ is the standard normal cumulative function. Notice that the crucial piece of information for the ex ante evaluation is that we have a measure of the cost of attending school, in this case, the opportunity cost of time, that is, the wage. Ex ante evaluation would be impossible if school-age children were precluded from engaging in market work or if the child wage did not vary.

A similar result holds if data were available for period one. The analogous decision rule for period one is

$$(5) \qquad s_1 = 1 \quad \text{iff} \quad \frac{\varepsilon_1}{\sigma} \geq \frac{w - \delta^2\alpha}{\sigma}; s_1 = 0$$

otherwise.

Period one data identify σ and $\delta^2\alpha$, which is sufficient to estimate the subsidy effect in period one, $\Phi[(w - \delta^2\alpha)/\sigma] - \Phi[(w - \delta^2\alpha - \tau)/\sigma]$. Notice that the period one effect will differ from the period two effect as long as the discount factor is not equal to one.

With both period one and period two data, δ and α are identified. Some policies require full

[2] The example is drawn from Todd and Kenneth I. Wolpin (2006).

[3] If the child wage varies within a village, and we observe only accepted wages, that is, wages for those children whose parents send them to work, we would need additional assumptions to identify the parameters. See Todd and Wolpin (2006).

identification. For example, suppose the subsidy took the form of a payment, τ^*, in period three based on the number of periods the child attended school. In that case, third-period utility would be $C_3 + (\alpha + t^*)S_3$. The decision rule in period two would now depend on the quantity $(w - \delta\alpha - \delta\tau)/\sigma$, which requires an estimate of δ. Maximizing the two-period likelihood function provides estimates of all the parameters of the model.

Given the estimates, a researcher would perform a number of diagnostics. One type of diagnostic would be to perform tests of model fit. The model predicts the school attendance rate in each period as well as how the attendance rate in each period varies with the child wage. The model also predicts that the attendance rate will be invariant to parent income, and that the attendance rate in the first period should be unrelated to the attendance rate in the second. The extent of the quantitative differences in the model predictions and the data can be assessed by conducting formal chi-square tests. Another diagnostic is to test the restrictions the model places on the data. As noted, the model is overidentified with two periods of data. We can obtain a separate estimate of σ using data from each period. The restriction that the two estimates are the same can be tested. A rejection of the restriction implies that a model that allows the variance of the shock to differ in the two periods is a better fit to the data.

Suppose that the model, even with time varying σ, is determined not to provide a good fit to the data. In this case, the researcher may have little confidence in the ex ante policy predictions from the model. So, having chosen a poor model to take to the data, what would a researcher do in practice? Obviously, change the model. In what directions would the researcher change the model? Presumably, in the directions in which the model poorly fit the data. Even in the stylized world of this example, there are many ways to change the model to improve its fit, and overfitting a model may adversely affect the ex ante policy prediction. Moreover, given that we are in the realm of "data mining," albeit in the context of estimating a fully parameterized behavioral model (what one might call structural data mining), the usual formal methods of model selection are no longer applicable.

Faced with poor model fit, a researcher might change the functional form. In this case, for example, if parent income and school attendance are related in the data, the utility function in period three can be augmented to include an interaction term between consumption and completed schooling (C_3S_3) or to allow for consumption to enter nonlinearly. A researcher may also add state variables to the problem. For example, if children who attend school in period one are more likely to attend in period two, period two utility may be augmented to include a term in period one school attendance, s_1. Alternatively, to capture the intertemporal pattern in attendance, one might allow for permanent unobserved heterogeneity in preferences for completed schooling. To fit overall attendance rates better, a researcher might add observable preference shifters, such as allowing preferences for completed schooling to depend on parent education. Finally, the researcher may make a different distributional assumption. In choosing from among this set, the researcher needs to be cognizant of identification and of overfitting (fitting the model to sample idiosyncrasies).

These possibilities are greatly magnified given that the actual world is considerably more complex than the stylized one presented above. For example, there are more than two decision periods, parent income fluctuates over time, child wages fluctuate over time and among children in the same village, households have more than one child, children are of more than one gender, and some children neither attend school nor work. The researcher must decide which of these are important to take into account in a model for which the purpose is to quantitatively evaluate the effect of the school attendance subsidy, and how to take them into account.

I would argue that there is no alternative to this type of data mining. Conducting specification tests is unconvincing because the model is by necessity the result of repeated pretesting. Estimating a fixed set of models and employing a model selection criterion (AIC, BIC) is equally unconvincing because models that result from repeated pretesting will tend to be very similar in terms of model fit. Given that view, how are we to choose among models?

II. A Pragmatic Approach to Model Selection

In Michael P. Keane and Wolpin (forthcoming), we have argued for a pragmatic view of empirical model building that recognizes there is

no "true" decision-theoretic model, only models that do better or worse in addressing particular questions. One criterion for model selection that fits within this view, as well as being consistent with current model-building practice, is to examine a model's out-of-sample predictive accuracy. If the purpose of the model is an ex ante policy evaluation, a more successful model would be one that makes a better prediction of the policy effect. But, how can we know which model will perform better prior to observing the policy?

Consider the Mexican program. Because of its experimental design, a researcher studying that program has available data on treated and control households. A comparison of school attendance rates of children in the treated and control villages provides an estimate of the program's impact. The social experiment was conducted using the same single subsidy schedule, which varied by grade level and child gender, that was used in the program's full implementation. Thus, the experiment was limited in the evaluations that could be performed. For example, the experiment could not inform the policymaker about what would have happened had the subsidy levels been doubled or halved, if instead of rewarding school attendance, the program had provided graduation bonuses, or if households had simply received income transfers. To answer those questions, one would need to build a model (or have conducted many more social experiments).

In selecting among models that would be capable of providing these additional ex ante policy evaluations, the question arises as how best to use the data. To address that issue, imagine a policymaker developing an evaluation strategy. The policymaker decides to select N researchers, each of whose task is to develop a model with which to perform the ex ante evaluation. The policymaker has two choices in terms of the data that can be provided to the researchers. One alternative is to give the researcher all the data, the pre- and post-program data on the control and treatment households. The researcher would then be able to use the variation in the direct cost of schooling provided by the experiment in order to estimate the model. The alternative is to hold out the post-program treatment households, in which case the researcher would have to rely only on child wage variation to identify the model.

The question is whether there is any gain to holding out the post-program data of the treated households, that is, to forgo using all of the variation in the data. One might argue that we would have more confidence in a model that best forecasted out of sample, particularly if the forecast were along the same policy dimension as the intended ex ante policy evaluation, than in a model that performed best within sample along that same policy dimension. The notion is that having been successful in extrapolating outside the support of the policy, that is, successfully predicting the post-program response of the treated group to the actual program, provides evidence that further extrapolation (double the subsidy, for example) also will be successful. Although there may be a sense in which such an assertion seems reasonable, I am aware of no formal justification for it.[4] Indeed, in terms of model selection, holding out a part of the sample would have to have an advantage in order to justify the information loss from estimating the model on a smaller sample with less variation.

The justification for holding out data is a pragmatic one. The necessary practice of structural data mining complicates the use of formal model testing procedures and yields models that will tend to look observationally equivalent in terms of within-sample model fit, although, strictly speaking, they may not be observationally equivalent. Out-of-sample forecasts that have not been contaminated by pretesting are more likely to diverge, due to sampling variation and to overfitting.

Social experiments provide an ideal setting within which to conduct out-of-sample model selection. Regime shifts provide an additional source of holdout samples. Keane and Wolpin (forthcoming) review a number of examples found in the literature. These papers make use of what is, from the researcher's perspective, a fortuitous event. Their common and essential element is the existence of some change radical enough to provide a degree of distance between the estimation and validation samples. The larger the change, the less likely are predicted and actual behavior in the validation sample to be close purely by chance. Waiting for social experiments or regime shifts to arise, given

[4] The idea seems to be a rather old one in psychology. Charles I. Mosier (1951) suggested the procedure, naming it "validity generalization," that is, validation by generalizing beyond the sample.

their rarity, does not lead to a viable research approach to model selection, however.

Recently, Keane and Wolpin (forthcoming) have proposed holding out a part of the sample that has the characteristics of a regime change. This "nonrandom holdout sample" would then be used for model selection. In their application, the hold-out sample takes advantage of the wide variation across US states that has existed in welfare policy. Specifically, they formulate and estimate a dynamic programming model of the joint schooling, welfare take-up, work, fertility, and marriage decisions of women using data from one group of US states (the estimation or "control" sample) and forecast these same decisions on another state (the validation or "treatment" sample) that differs dramatically in the generosity of its welfare program. The holdout sample meets the criterion of a regime shift, namely that the "new" policy is far outside the support of current policy variation that was used for estimation.

Keane and Wolpin (forthcoming) argued that choosing a nonrandom holdout sample is a viable strategy for model selection and one that researchers can implement in many settings. This strategy is particularly applicable in policy evaluations that rely on existing policy variation. Examples in which holdout samples are not direct policy variables might be: (a) a researcher who wishes to perform an ex ante evaluation of the effect of potential social security reforms on retirement behavior might think of choosing a control or estimation sample that included individuals without private pensions and a treatment or holdout sample that included individuals with private pensions; or (b) a researcher who wished to perform an ex ante evaluation of the effect of potential child care subsidy programs on a female labor supply model might choose estimation and holdout samples that differed in their numbers of children.[5]

III. Conclusion

Given the many assumptions that researchers make in order to carry out an ex ante policy evaluation and the necessarily extensive role that pretesting plays in building models, it is not a viable research strategy to check the robustness of models to individual assumptions or to select among models based on within-sample fit. Researchers, beginning with the same question and using the same data, will generally differ along many dimensions in the modeling assumptions they make, and resulting models will tend to be indistinguishable in terms of model fit. Having a generally accepted criterion for judging the relative success of models would seem to be a useful step toward choosing among models. Assessing a model's predictive accuracy for a nonrandomly selected holdout sample, one that is relevant to the research question addressed, is one such criterion.[6]

[5] Using a holdout sample for model selection discards information that would be useful in estimation. To increase precision, models can be reestimated incorporating the data from the holdout sample. Ex ante policy evaluation can then be carried out using these new estimates.

[6] Keane and Wolpin (2006) also find that, as with any procedure, choosing a model by strictly adhering to an out-of-sample prediction criterion can be misleading. In their example, the best out-of-sample prediction model led to highly implausible policy effects.

REFERENCES

Keane, Michael P., and Kenneth I. Wolpin. 2006. "The Role of Labor and Marriage Markets, Preference Heterogeneity and the Welfare System in the Life Cycle Decisions of Black, Hispanic and White Women." Unpublished.

Keane, Michael P., and Kenneth I. Wolpin. Forthcoming. "Exploring the Usefulness of a Non-Random Holdout Sample for Model Validation: Welfare Effects on Female Behavior." *International Economic Review.*

Marschak, Jacob. 1953. "Economic Measurements for Policy and Prediction." In *Studies in Econometric Method (Cowles Commission for Research in Economics Monograph No. 14)*, ed. William C. Hood and Tjalling C. Koopmans, 1–26. New York: Wiley.

Mosier, Charles I. 1951. "Problems and Designs of Cross-Validation." *Educational and Psychological Measurement*, 11(1): 5–11.

Todd, Petra. 2006. "Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated." Unpublished.

Todd, Petra B., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico." *American Economic Review*, 96(5): 1384–1417.