Research in Economics 70 (2016) 332–345

Contents lists available at ScienceDirect

Research in Economics

journal homepage: www.elsevier.com/locate/rie

To hold out or not to hold out

Frank Schorfheide [a,*], Kenneth I. Wolpin [b]

[a] University of Pennsylvania, Department of Economics, 3718 Locust Walk, Philadelphia, PA 19104-6297, United States
[b] Rice University, Department of Economics – MS22, P.O. Box 1892, Houston, TX 77251-1892, United States

ARTICLE INFO

Article history:
Received 11 February 2016
Accepted 18 February 2016
Available online 10 March 2016

Keywords:
Bayesian analysis
Model selection
Principal–agent models
Randomized controlled trials

ABSTRACT

A recent literature has developed that combines two prominent empirical approaches to ex ante policy evaluation: randomized controlled trials (RCT) and structural estimation. The RCT provides a "gold-standard" estimate of a particular treatment, but only of that treatment. Structural estimation provides the capability to extrapolate beyond the experimental treatment, but is based on untestable assumptions and is subject to structural data mining. Combining the approaches by holding out from the structural estimation exercise either the treatment or control sample allows for external validation of the underlying behavioral model. Although intuitively appealing, this holdout methodology is not well grounded. For instance, it is easy to show that it is suboptimal from a Bayesian perspective. Using a stylized representation of a randomized controlled trial, we provide a formal rationale for the use of a holdout sample in an environment in which data mining poses an impediment to the implementation of the ideal Bayesian analysis and a numerical illustration of the potential benefits of holdout samples.

1. Introduction

A recent literature has developed that combines two prominent empirical approaches to ex ante policy evaluation: randomized controlled trials (RCT) and structural estimation (see, for example, Wise, 1985, Todd and Wolpin, 2006, or Duflo et al., 2012). The RCT provides a "gold-standard" estimate of a particular treatment, but only of that treatment. Structural estimation provides the capability to extrapolate beyond the experimental treatment, but is based on untestable assumptions and is subject to structural data mining. Combining the approaches by holding out from the structural estimation exercise either the treatment or control sample (or a fraction of both) allows for external validation of the underlying behavioral model. Although having intuitive appeal, the use of holdout samples is methodologically not well grounded. For instance, Bayesian analysis prescribes using the *entire* sample to form posterior model probabilities and using the resulting predictive distributions to characterize policy effects.

The contributions of this paper are twofold. First, we provide a formal, albeit stylized, framework in which Bayesian inference and decision-making is optimal but data mining poses an impediment to the implementation of the ideal Bayesian solution. Throughout this paper we use the term data mining to label a process by which a modeler tries to improve the fit of the model during estimation, for example, through changing functional forms, adding observed or latent state variables, etc. Second, we provide a numerical illustration of the potential costs of data mining and the potential benefits of holdout

* Corresponding author.
E-mail addresses: schorf@ssc.upenn.edu (F. Schorfheide), kenneth.i.wolpin@rice.edu (K.I. Wolpin).

http://dx.doi.org/10.1016/j.rie.2016.02.001
1090-9443/

samples that are designed to discourage data mining. Losses are measured relative to the optimal Bayesian decision. Our illustration implies that holdout samples can provide a basis for assessing the relative credibility of competing models.

It is important to emphasize that our paper does not argue the well-established point that measures of in-sample goodness-of-fit need to be explicitly (e.g., Schwarz, 1978's Bayesian information criterion) or implicitly (e.g., Stone, 1977's cross-validation approach) adjusted for model complexity to avoid overfitting and enable consistent or efficient model selection. This we take for granted. Our analysis will show that measures of model fit that are penalized for model dimensionality can also be misleading if the modeler has access to the full sample. This will be the case if the modeler has engaged in a sequence of data-based modifications of the structural model and reports a measure of penalized model fit only for the final specification that is the outcome of a data-mining process. Only the validation based on a holdout sample can discourage (undesirable features of) data mining and unduly optimistic assessments of model fit. Although data mining is often informally cited as an argument in favor of the use of holdout samples, to the best of our knowledge our paper is the first to provide a formalization. The use of holdout samples to discourage data mining extends well beyond the RCT and program-evaluation literature and is widespread in the social sciences. For instance, Schorfheide and Wolpin (2012) discuss examples from time series analysis and macroeconomic modeling. In the psychology literature, Mosier (1951) suggested the use of holdout samples, naming it "validity generalization," that is, validation by generalizing beyond the sample.

Our framework can be viewed as a principal–agent setup. A policy maker is the principal, who would like to predict the effects of a treatment at varying treatment levels. The policy maker has access to data from a social experiment, conducted for a single treatment level. To assess the impact of alternative treatments, the policy maker engages two structural modelers, the agents, each of whom estimates their structural model and provides measures of predictive fit.[1] We assume that the modelers are rewarded in terms of the fit of their model. Two mechanisms are considered. Under the *no-holdout* mechanism, the modelers have access to the full sample of observations and are evaluated based on the so-called marginal likelihood functions that they report. In a Bayesian framework, marginal likelihoods are used to update model probabilities. Because the modelers have access to the full sample, there is an incentive to modify their model specifications and to overstate the marginal likelihood values. We refer to this behavior as data mining.

Under the *holdout* mechanism, on the other hand, the modelers have access only to a subset of observations and are asked by the policy maker to predict features of the sample that is held out for model evaluation. Building on an old result by Winkler (1969) on log scoring rules, the holdout mechanism is designed so that the modelers truthfully reveal their subjective beliefs about the holdout sample. However, predictive distributions for the holdout sample are not as informative as marginal likelihoods for the entire sample, which is why the policy maker is unable to implement the full Bayesian analysis with this mechanism.

Our analysis abstracts from the complexities of dynamic optimization-based structural models that are used in empirical applications. To capture the essence of a complex problem in a stylized framework, we equip each of the modelers with a linear regression model. These models are structural in the sense that we impose a cross-coefficient restriction that allows the modelers to identify the parameter that controls the magnitude of the treatment effect solely based on variation in an exogenous regressor. At first glance, this assumption may appear overly restrictive. An example in which identification might be problematic is where treatment provides an intrinsic benefit or cost, for example, a stigma effect in the case of a welfare program.[2] However, by varying the variance of the exogenous regressors relative to the magnitude of the treatment in the RCT, we can capture the fact that estimates of the treatment effect based solely on variation in the exogenous regressor may be very imprecise in comparison to estimates that also utilize information from outcome differentials among individuals in treatment and control groups.[3]

We represent the act of data mining as data-based modifications of the prior distributions that the modelers use to obtain posteriors. The modified prior distributions relax the cross-coefficient restrictions in an attempt to fit the treatment effect. In the context of actual structural modeling, this modification of the prior is meant to capture functional form adjustments of agents' preferences and firms' production functions, or the inclusion of additional heterogeneity, to match the treatment effect in the data.

To keep the analysis manageable and transparent, our framework does not account for a model development phase. We equip each agent with a regression model, but do not attempt to explain how the modelers arrived at their regression specifications. The outcome of the model development stage is typically that researchers have arrived at specifications that are difficult to distinguish based on the available data. This is captured in our framework by focusing on parameterizations that imply that both models have non-trivial posterior probabilities and that the frequency (in repeated sampling) with which the highest posterior probability model equals the "true" model is clearly less than one.

Although we are able to give a qualitative characterization of the behavior of the modelers under the two mechanisms based on analytical derivations, we use a numerical example to illustrate how the size and the composition (in terms of observations from the control and treatment groups) of the holdout sample affects the risk of the policy maker. We find that the *holdout* mechanism dominates the *no-holdout* mechanism because of the data mining that occurs if the modelers have access to the full sample. The lowest level of risk is attained by holding back 50% of the sample (where the control and

[1] A structural model is one in which parameters are policy (treatment) invariant.

[2] See Attanasio et al. (2012) and Wolpin (2013) for further discussion of this point.

[3] Ferrall (2012), in a study of the Canadian Self-sufficiency Project, finds that using only the control group in estimation leads to much less precise parameter estimates.