RISE | RICE INITIATIVE *for the* STUDY *of* ECONOMICS

RISE Working Paper 14-018

"To Hold Out or Not To Hold Out"

by

Frank Schorfheide, Kenneth I. Wolpin

RICE
Department of Economics
Baker Hall, MS22
6100 Main Street, Houston, Texas 77005
https://economics.rice.edu

To Hold Out or Not To Hold Out*

Frank Schorfheide† Kenneth I. Wolpin‡

July 17, 2014

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 the 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 illiustration of the potential benefits of holdout samples. (JEL C11, C31,C52)

*Correspondence: University of Pennsylvania, Department of Economics, 3718 Locust Walk, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu (F. Schorfheide);kenneth.i.wolpin@rice.edu or wolpink@ssc.upenn.edu (K. Wolpin). We thank Frank Diebold, Michael Keane, George Mailath, Chris Sims, Tao Zha as welll as seminar participants at the 2012 AEA meetings, Columbia, Princeton, USC and the University of Pennsylvania for helpful comments and suggestions.

†University of Pennsylvania

‡Rice University and University of Pennsylvania

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, Hanna, and Ryan (2011)). 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. Data mining in this context is 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 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,

has engaged in a sequence of data-based modifications of his 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

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

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 the modelers each 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

[2]See Attanasio, Meghir, and Santiago (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.

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 treatment sample are of equal size) and providing the modelers only with data either from the control or from the treatment group.

Conditional on it being desirable to provide the modelers with only a fraction of the available data, our result about the optimal composition of the estimation sample may appear counterintuitive. After all, including observations from both the control and treatment group will allow the modelers to generate sharper estimates of the treatment effect. At the same time, this additional information will make the treatment effect estimates more similar across models, which makes it more difficult for the policy maker to determine the highest-posterior-probability model based on the limited information from the predictions for the holdout sample.

Our paper is related to several branches of the economics literature. We draw on the literature on scoring rules and the evaluation of probability assessors when setting up the payoff scheme for the modelers. Winkler (1969) showed that log predictive densities create the incentive to truthfully reveal subjective probabilities. Further results on the evaluation of probability assessors can be found in the textbook by Bernardo and Smith (1994) and in the literature on testing of experts, e.g. Sandroni (2003). Our setup assumes that the payoff to Modeler 1 does not depend on the predictions made by Modeler 2 (and vice versa). Thus, by assumption we ignore possible strategic interactions among the modelers, which are the subject of the literature on incentives of macroeconomic forecasters, e.g. Laster, Bennet, and Geoum (1999) and Lamont (2002).

Our work is also related to the literature on pooling of probability distributions and the so-called expert problem, e.g., French (1985), Lindley (1985), Genest and Zidek (1986), and Clemen and Winkler (2007). In this literature, a decision maker seeks advice from a panel of experts and has to aggregate opinions expressed in terms of probability distributions. However, our setup differs from this literature in that our policy maker can control the information that is made available to the modelers (experts) and that the experts may engage in a particular form of data mining.

Leamer (1978) studies the effect of specification searches on inference in non-experimental settings. Lo and MacKinlay (1990) and White (2000) provide methods of correcting statistical inference procedures for so-called data snooping. An example of data snooping is to run many preliminary regressions based on a large set of explanatory variables, but only reporting results based on a specification in which a regressor appeared to be significant and able to, say, predict stock returns. This literature has focused on correcting standard error estimates for data snooping. Our concept of data mining is somewhat different from the act of searching among a large pool of regressors. We focus on data-based modifications of structural economic models, e.g. changing functional forms, that are designed to improve in-sample fit.

Holdout samples play an important role in cross validation approaches, e.g. Stone (1977). The cross-validation literature showed that model validation on pseudo-holdout samples can generate a measure of fit that penalizes model complexity. In our paper, however, the goal is not to generate a new penalty term for in-sample fit of an econometric model. In fact, the marginal likelihoods that are used in a Bayesian framework to construct posterior model probabilities and serve as a benchmark for our analysis, can be interpreted as maximized likelihood functions that are penalized for the number of free parameters in the model.

The remainder of this paper is organized as follows. For concreteness, in Section 2 we describe a working example in which a policy maker is trying to determine the optimal level of a school-attendance subsidy. Using a number of simplifying assumptions, we are able to represent the structural models for the analysis of the policy question by simple univariate linear regressions. The Bayesian solution to predicting the effects of a school-attendance

subsidy is presented in Section 3. Section 4 contains the principal-agent setup that is used to capture the potential benefits of weighting (or selecting) among structural models based on predictions for holdout samples and Section 5 provides the numerical illustration. Finally, we conclude in Section 6.

2 A Working Example

To analyze the potential benefits of holdout samples we consider the problem of evaluating the impact of a monetary subsidy to low-income households based on school attendance of their children. It is assumed that prior to the policy change there is no direct tuition cost of schooling.[4] The goal is to determine an optimal level of the subsidy that trades off the costs of the subsidy program with its effect on the attendance rate. A social experiment is conducted in which a randomly selected treatment sample is offered a school subsidy at the level $s = \bar{s}$, whereas no subsidy is provided to the households in the control sample, that is, $s = 0$. The outcome variable for household i, $i = 1, \ldots, n$, is denoted by h_i and is continuous, e.g. attendance measured in hours.

Because, in practice, it is too costly to make the treatment sample sufficiently large to allow the treatment to be measured at a variety of subsidy levels, the policy maker has to rely on the estimation of structural models to extrapolate the treatment effect to other levels of treatment $s^* \neq \bar{s}$. We assume that there are two such structural models M_j, $j = 1, 2$. Each household i solves the following optimization problem to determine the number of hours to send their child to school:[5]

$$\max_{c \in \mathbb{R}^+, h \in [0,1]} U_j(c, h; z, u, \vartheta_j) \quad \text{s.t. } c = inc + w(T - h) \tag{1}$$

Here $U_j(\cdot)$ is a model-specific utility function, parameterized in terms of ϑ_j, c is household consumption, $h \in [0, T]$ is hours spent in school, where T is the total endowment of time,

[4] Tuition cost variation permits the estimation of the effect of introducing a subsidy nonparametrically for subsidy levels for which net tuition is within the domain of the tuition variation, Ichimura and Taber (2000).

[5] This example is taken from Todd and Wolpin (2008).

z is a vector of observable household characteristics, u is a random variable that captures unobservable preference heterogeneity, and inc is parental income.

We denote the optimal attendance decision by $h = \varphi_j(inc, w; z, u, \vartheta_j)$. An attendance subsidy s modifies the households' budget constraint to

$$c = inc + w(T-h) + sh = (inc + s) + (w - s)(T - h) = \widetilde{inc} + \tilde{w}(T - h). \tag{2}$$

The optimal attendance choice in the presence of a subsidy is

$$h^* = \varphi_j(\widetilde{inc}, \tilde{w}; z, u, \vartheta_j). \tag{3}$$

The modified budget constraint (2) implies that variation in household income and wage w are sufficient to identify the effect of a school subsidy on attendance (Todd and Wolpin (2008)). In fact, it is a key feature of many structural models that the parameters necessary for a counterfactual policy analysis can be identified even if the sample contains no variation in the policy instrument.

In order to simplify the subsequent exposition, suppose that the decision rule (3) is linearized and represented in the following stylized form, where hours h is replaced by y and x is a scalar characteristic of the household (replacing inc, w, and z):

$$y_i = x_{i,j}\beta_j + s_i\theta + u_i \quad u_i|(x_i, s_i) \sim iidN(0, 1). \tag{4}$$

For expositional convenience, we set the variance of u_i equal to one. The j subscripts in (4) capture the different assumption embodied in the two models about the relevant characteristic x that affects the outcome. As previously mentioned, an important feature of structural models is that they contain restrictions that allow the identification of policy effects without sample variation in the policy instrument. To capture this aspect in our regression model (4), we impose the restriction $\theta = \beta_j$.[6] Thus, variation in $x_{i,j}$ is sufficient to obtain an estimate of the subsidy effect.[7] Because we will subsequently use matrix notation, let X_j be

[6] In the example, if there is no income effect on school attendance, e.g., if the utility function is quasi-linear in consumption and if the utility function is quadratic in hours of school attendance, then x would be the child wage and $\theta = -\beta_j$.

[7] For convenience we assume away the existence of heterogeneous treatment effects.

the $n \times 1$ vectors with elements $x_{i,j}$, $X = [X_1, X_2]$, and let Y and S be the $n \times 1$ vectors with elements y_i and s_i, respectively. For notational convenience we drop the intercept from (4). However, in the numerical illustration in Section 5 we demean all regressors on the respective subsamples considered for estimation, which essentially introduces an intercept.

3 First-Best Analysis

We proceed by specifying a formal decision problem for a policy maker who has to predict the effect of a subsidy that differs in magnitude from the subsidy considered in the RCT. We deliberately use a setup in which Bayesian analysis can provide the optimal (or first-best) decision. In Section 4 below we will introduce a friction that provides an impediment to the implementation of the optimal decision. The Bayesian approach requires us to specify prior distributions for the parameters of models M_1 and M_2 as well as prior probabilities for the models themselves. Both models are equipped with the prior distribution $\theta \sim N\big(0, 1/\lambda^2\big)$. The density of this prior is denoted by $p(\theta|M_j)$. Overall, this leads to

$$M_j: \quad Y = \tilde{X}_j\theta + U, \quad U|(X,S) \sim N(0,I), \quad \theta \sim N\left(0, \frac{1}{\lambda^2}\right), \quad j = 1,2, \tag{5}$$

where $\tilde{X}_j = X_j + S$ and the variance of U is assumed to be the identity matrix I for analytical convenience. Given the randomization of the RCT, the selection of the treatment group is independent of the observable characteristics, that is, $p(X,S) = p(X)p(S)$. We assume that the marginal densities of $p(X)$ and $p(S)$ do not depend on θ and are the same for both models. These assumptions have the convenient implication that $p(X,S)$ cancels from most of the formulas presented below and we can base our derivations on the model-specific distributions of $Y|(X,S)$. The prior model probabilities assigned to models M_1 and M_2 are denoted by $\pi_{j,0} = 1/2$, $j = 1,2$.

The overall posterior distribution of the treatment effect is given by the mixture

$$p(\theta|Y,X,S) = \sum_{j=1,2} \pi_{j,n} p(\theta|Y,X,S,M_j), \tag{6}$$

where

$$\pi_{j,n} = \frac{\pi_{j,0}p(Y|X,S,M_j)}{p(Y|X,S)}, \quad p(Y|X,S) = \sum_{j=1,2} \pi_{j,0} p(Y|X,S,M_j). \tag{7}$$

Here $p(\theta|Y,X,S,M_j)$ is the posterior density of θ conditional on model M_j, $\pi_{j,n}$ is the posterior probability of model M_j, $p(Y|X,S,M_j)$ is the marginal likelihood of M_j, and $p(Y|X,S)$ is the marginal likelihood of the mixture of M_1 and M_2.

We assume that the policy maker's goal is to predict the outcome for an individual that receives a subsidy $s^* \neq s$ and has characteristics (x_1^*, x_2^*) and $u^* = 0$.[8] The predictor of y is denoted by $\hat{y}$ and evaluated under a quadratic loss function. The posterior expected loss (we will subsequently refer to expected loss as *risk*) associated with this decision is given by

$$\rho(\hat{y}|Y,X,S) = \sum_{j=1,2} \pi_{j,n} \int_\theta \left((x_j^* + s^*)\theta - \hat{y}\right)^2 p(\theta|Y,X,S,M_j)d\theta. \tag{8}$$

Under the quadratic loss function the optimal predictor is given by the posterior mean of the outcome:

$$\hat{y}^* = \sum_{j=1,2} \pi_{j,n}(x_j^* + s^*) \int \theta p(\theta|Y,X,S,M_j)d\theta, \tag{9}$$

which is a weighted average of the posterior mean predictions of M_1 and M_2. Moreover, we can decompose the posterior risk of alternative predictors into

$$\rho(\hat{y}|Y,X,S) = \rho(\hat{y}^*|Y,X,S) + (\hat{y} - \hat{y}^*)^2. \tag{10}$$

This decomposition makes clear that the Bayes predictor $\hat{y}^*$ is first-best and any alternative predictor that deviates from the Bayes predictor is suboptimal. The posterior risk conditions on the observations (Y,X,S). In the numerical illustration in Section 5 We focus on the integrated risk, which averages over (Y,X,S):

$$\begin{aligned} \mathcal{R}(\hat{y}) &= \int_{Y,X,S} \rho(\hat{y}^*|Y,X,S)p(Y|X,S)p(X,S)d(Y,X,S) + \Delta(\hat{y}) \\ \Delta(\hat{y}) &= \int_{Y,X,S} (\hat{y} - \hat{y}^*)^2 p(Y|X,S)p(X,S)d(Y,X,S). \end{aligned} \tag{11}$$

Equation (11) highlights that the integrated risk is also minimized by the Bayes predictor $\hat{y}^*$. We refer to $\Delta(\hat{y})$ as the (integrated) risk differential.

To calculate the optimal predictor in (9) we need to evaluate $p(\theta|Y,X,S,M_j)$ and $\pi_{j,n}$. The model-specific posterior for θ, the treatment effect, is given by

$$p(\theta|Y,X,S,M_j) = \frac{p(Y|X,S,\theta,M_j)p(\theta|M_j)}{p(Y|X,S,M_j)}, \tag{12}$$

[8] Because our model is linear and $\mathbb{E}[u^*] = 0$, the assumption that $u^* = 0$ is inconsequential. However, it simplifies the notation a bit.

where $p(\theta|M_j)$ is the prior distribution of θ under model M_j. The model specification in (5) implies that this posterior distribution takes the form

$$\theta|(Y,X,S,M_j) \sim N\left(\hat{\theta}_j^*, (\lambda^2 + \tilde{X}_j'\tilde{X}_j)^{-1}\right), \quad \hat{\theta}_j^* = (\tilde{X}_j'\tilde{X}_j + \lambda^2)^{-1}\tilde{X}_j'Y \tag{13}$$

The posterior model probabilities $\pi_{j,n}$ are a function of the marginal likelihoods (see (7)), which in the linear Gaussian regression model can be calculated analytically and take the form

$$\begin{aligned} p(Y|X,S,M_j) &= (2\pi)^{-n/2}|1+\tilde{X}_j'\tilde{X}_j/\lambda^2|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[Y'(I-\tilde{X}_j(\tilde{X}_j'\tilde{X}_j+\lambda^2)^{-1}\tilde{X}_j')Y]\right\}. \end{aligned} \tag{14}$$

The exponential term captures the goodness of in-sample fit, whereas the term $|1+\tilde{X}_j'\tilde{X}_j/\lambda^2|^{-1/2}$ can be interpreted as a penalty for model complexity. The larger λ, and thus the less diffuse and more restrictive is the prior distribution, the less complex is the model. In fact, for $\lambda = \infty$, there is no free parameter to be estimated. On the other hand, a more variable regressor makes the model appear more complex. It requires a smaller value of λ and thus the prior is in relative terms more diffuse.

4 A Principal-Agent Problem

We now introduce a friction that makes the implementation of the first-best Bayesian analysis described in the previous section infeasible. We assume that the computation of model conditional posteriors $p(\theta|Y,X,S,M_j)$ in (13) and marginal likelihoods $p(Y|X,S,M_j)$ in (14) is executed by two expert modelers (agents). In a second stage, a policy maker (principal) aggregates the results that he obtains from the modelers to form a prediction $\hat{y}$. In some applications this assumption might be literally satisfied in the sense that a government agency conducts the social experiment and hires academic consultants to provide an analysis of the policy effects. In other instances, the policy maker might correspond to the economics profession at large as it is investigating the effectiveness of social programs and the agents correspond to economists who conduct research on the effects of a particular policy. Because

different individuals are assumed to be involved in the two stages of the analysis, incentive problems potentially arise. These incentive problems, in turn, can provide a rationale for holdout samples.

We proceed by describing the objective and constraints of the policy maker in Section 4.1. We then discuss two mechanisms that the policy maker could use to set incentives for the modelers in Section 4.2. One of the mechanisms involves a holdout sample. In the other mechanism, the modelers have access to the full data set. In Section 4.3, we characterize two options that are available to the modelers: (i) Bayesian analysis of model M_j based on the data provided by the policy maker and (ii) in-sample data mining, which is represented by a modification of the prior distribution. Finally, we discuss the optimal choices of the modelers under the two mechanisms in Section 4.4.

4.1 The Policy Maker

We assume that the social experiment described in Section 2 is conducted by a policy maker. The policy maker has access to all the data from the experiment, but can estimate only the treatment effect in the experiment, namely by taking the difference in means between the treatment and the control group. The estimator of the treatment effect can be represented as coming from the statistical model

$$M_{pm}: \quad Y = S\theta + V, \tag{15}$$

where V is a $n \times 1$ vector of error terms and we use the pm subscript to denote policy maker. The resulting estimator of the treatment effect is

$$\hat{\theta}_{pm} = (S'S)^{-1}S'Y. \tag{16}$$

Given that the subsidy takes on only two values, $s = 0$ or $s = \bar{s}$, the policy maker's statistical model M_p cannot be used to extrapolate the treatment effect to other levels of treatment $s \neq \bar{s}$. For that purpose, the policy maker engages the two modelers to analyze their structural models M_1 and M_2. His objective is to obtain a predictor that minimizes the integrated risk $\mathcal{R}(\hat{y})$ in (11). Thus, ideally, the policy maker would like to reproduce the Bayesian prediction $\hat{y}^*$.

4.2 Mechanisms Available to the Policy Maker

We consider two potential mechanisms that the policy maker can use to obtain the decision-relevant information from the modelers. Under the first mechanism, the modelers receive the entire sample (Y, X, S). Under the second mechanism, the policy maker splits the sample and hands the modelers only a subset of the observations. The policy maker has discretion about the size of the holdout sample and its composition in terms of observations from the treatment and control group.

No-Holdout Mechanism. The policy maker gives the modelers access to the entire data set (Y, X, S). In turn, they are asked to report a marginal data density $\tilde{p}_j(Y|X, S)$ and a posterior distribution for the treatment effect $\tilde{p}_j(\theta|Y, X, S)$. We use $\tilde{p}(\cdot)$ rather than $p(\cdot)$ to allow for the possibility that the modelers do not truthfully reveal these two objects. Only if the reported densities coincide with the actual densities in (13) and (14) can the policy maker implement the first-best Bayesian decision. We assume that the compensation of the modelers is a function of how well their models are able to fit the data, adjusting for model complexity.[9] More specifically, under the *no-holdout* (NH) mechanism the payoff is equal to the reported log marginal likelihood

$$\Pi_{NH}\big(\tilde{p}_j(Y|X, S)\big) = \ln \tilde{p}_j(Y|X, S). \tag{17}$$

The policy maker updates the model probabilities according to

$$\tilde{\pi}_{j,n} = \frac{\pi_{j,0}\tilde{p}_j(Y|X, S)}{\pi_{1,0}\tilde{p}_1(Y|X, S) + \pi_{2,0}\tilde{p}_2(Y|X, S)}. \tag{18}$$

Note that the payoff for Modeler 1 is independent of the action taken by Modeler 2, and vice versa. Thus, we abstract from strategic interactions between the modelers.

Holdout Mechanism. The modelers receive the full sample of covariates and treatment levels (X, S), but only a subset of the outcome data Y from the policy maker. The outcome data are partitioned into $Y' = [Y_r', Y_p']$, where Y_r is a *regression* sample that is given to the

[9] We rule out any kind of collusive behavior among the modelers including the aggregation of model predictions and the redistribution of payoffs.

modelers for estimation purposes and Y_p is a *holdout* or *prediction* sample that can be used by the policy maker to evaluate predictions.[10]

We assume that the policy maker updates the model probabilities based on the predictive densities $\tilde{p}_j(\hat{\theta}_{pm}|Y_r, X, S)$ for the difference-in-means estimate of the treatment effect instead of the predictive density $\tilde{p}_j(Y_p|Y_r, X, S)$ for the entire holdout sample. In realistic applications the precise evaluation of $p(Y_p|Y_r, X, S, M_j)$ for one particular sample is often challenging and time consuming. Computing this density for all possible realizations Y_p is a daunting task. The difference-in-means estimate $\hat{\theta}_{pm}$, on the other hand, is a univariate statistic in our application and reporting a predictive density is straightforward. It could easily be graphed or tabulated. In sum, while the use of a density for Y_p is theoretically attractive, it is difficult, if not infeasible to implement. The current practice in the treatment-effect literature comes closest to choosing model probabilities based on the $\hat{\theta}$-predictive density, as for example in Todd and Wolpin (2006) and Duflo, Hanna, and Ryan (2011).

The mechanism unfolds in two stages. First, the policy maker asks the modelers to provide a predictive density $\tilde{p}_j(\hat{\theta}_{pm}|Y_r, X, S)$ for their estimate of the treatment effect given by (16). Similar to the compensation under the *no-holdout* mechanism, we assume that under the *holdout* (H) mechanism the compensation takes the form:

$$\Pi_H\big(\tilde{p}_j(\hat{\theta}_{pm}|Y_r, X, S)\big) = \ln \tilde{p}_j(\hat{\theta}_{pm}|Y_r, X, S). \tag{19}$$

The predictive densities $\tilde{p}_j(\hat{\theta}_{pm}|Y_r, X, S)$ are then used to update the model probabilities:

$$\tilde{\pi}_{j,n} = \frac{\pi_{j,0}\tilde{p}_j(\hat{\theta}_{pm}|Y_r, X, S)}{\pi_{1,0}\tilde{p}_1(\hat{\theta}_{pm}|Y_r, X, S) + \pi_{2,0}\tilde{p}_2(\hat{\theta}_{pm}|Y_r, X, S)}. \tag{20}$$

Second, once the model probabilities are updated the policy maker makes all the outcome data available and asks the modelers to re-estimate their models and report $\tilde{p}_j(\theta|Y, X, S)$. Allowing the modelers to re-estimate the parameters on the full sample avoids an unnecessary loss of information about θ that would put the mechanism at a clear disadvantage. After all, the rationale of holdout samples is merely to avoid distortions in model probabilities due to data-mining. We use $\tilde{p}_j(\theta|Y, X, S)$ to denote the posterior of θ reported by the modelers.

[10]For the modelers' inference it is inconsequential, given randomization, whether they have access to the full sample of regressors or just the subsample that corresponds to Y_r. We assumed the former because it simplifies the notation.

4.3 The Choice Set of the Modelers

We assume that the modelers can choose between the following two options: (i) do not engage in data mining and report results from the Bayesian analysis of M_j based on the sample provided by the policy maker; (ii) engage in a form of data-mining that uses the available sample to break the model-implied link between β_j and θ and shift the prior distribution for (β_j, θ) toward a region of the parameter space favored by the available data.

Option (i): Bayesian Analysis of M_j. Under the *no-holdout* mechanism the modelers have access to the full sample and report the marginal likelihood for Y, which is given in (14). Under the *holdout* mechanism the modelers can compute the predictive likelihood $p(Y_p|Y_r, X, S, M_j)$, which in turn implies a predictive density for $\hat{\theta}_{pm}(Y_p, Y_r)$, denoted by $p(\hat{\theta}_{pm}|Y_r, X, S, M_j)$. The corresponding full-sample posterior for θ is given by $p(\theta|Y, X, S, M_j)$. Under a quadratic loss function the point estimate of θ is the posterior mean.

Option (ii): In-Sample Data-Mining. We represent in-sample data mining as data-based modification of the prior distribution associated with model M_j. In addition to breaking the tight link between θ and β_j, this form of data mining also shifts the prior toward an area of the parameter space in which the likelihood function is relatively high. It is supposed to capture a practice whereby a researcher inspects the data and, depending on the properties of the data, decides which features (e.g., functional forms for utility and production functions, adjustment cost mechanisms, household or firm heterogeneity) to include in the model and which to leave out, without accounting for this specification search subsequently.

In our working example, given data from both the treatment and control samples, the data-mining prior is constructed as follows. We begin by breaking the link between θ and β_j by considering the unrestricted model

$$Y = X_j\beta_j + S\theta + U. \tag{21}$$

Let $Z_j = [X_j, S]$ and $\psi_j = [\beta_j, \theta]'$. Moreover, we assume that the modeler uses a prior distribution that is centered at the peak of the likelihood function (which in our model

coincides with the OLS estimator). Overall, the data-mined model takes the form

$$\begin{aligned}\tilde{M}_j \quad : \quad & Y = Z_j\psi_j + U, \quad \psi_j \sim N\left(\tilde{\psi}_j, (\kappa Z_j'Z_j)^{-1}\right) \\ & \tilde{\psi}_j = (Z_j'Z_j)^{-1}Z_j'Y.\end{aligned} \tag{22}$$

The parameter κ scales the prior precision of ψ_j. Based on model $\tilde{M}_j$ it is possible to compute either the marginal likelihood function $p(Y|X,S,\tilde{M}_j)$ or the predictive density $p(\hat{\theta}_{pm}|Y_r,X,S,\tilde{M}_j)$. The posterior distribution $p(\theta|Y,X,S,\tilde{M}_j)$ under the data-mined model remains normal, but it has a different mean and variance than the posterior in (13). Defining $M_{X_j} = I - X_j(X_j'X_j)^{-1}X_j'$ we obtain:

$$\theta|Y,X,S,\tilde{M}_j \sim N\left(\tilde{\theta}_j, \left((\kappa+1)S'M_{X_j}S\right)^{-1}\right), \quad \tilde{\theta}_j = (S'M_{X_j}X_j)^{-1}S'M_{X_j}Y. \tag{23}$$

Under a quadratic loss function the point estimate of θ associate with the data-mined model $\tilde{M}_j$ is $\tilde{\theta}_j$. Under *in-sample data-mining* the cross-equation restrictions that allow for a more efficient estimation of θ are abandoned and the the stated measure of uncertainty is severely distorted.

4.4 Optimal Choices of the Modelers

Having described the potential choices of the modelers, we can now discuss their actual choices in the *no-holdout* and the *holdout* mechanisms.

No-Holdout Mechanism. If modeler j chooses Option (i), i.e., he truthfully reports the results from the Bayesian analysis of M_j, then he is evaluated based on $p(Y|X,S,M_j)$. On the other hand, if modeler j chooses Option (ii), in-sample data mining, then he is rewarded based on $p(Y|X,S,\tilde{M}_j)$. In order to determine the optimal choice of the modeler we now compare the two marginal likelihood functions. The marginal likelihood associated with Option (i) is given in (14). The marginal likelihood function associated with Option (ii) takes the form

$$p(Y|X,S,\tilde{M}_j) = (2\pi)^{-n/2}|1/\kappa+1|^{-1/2}\exp\left\{-\frac{1}{2}[Y'(I - Z_j(Z_j'Z_j)^{-1}Z_j)Y]\right\}, \tag{24}$$

where $Y'(I - Z_j(Z_j'Z_j)^{-1}Z_j)Y$ is the sum of squared residuals (SSR) for the unrestricted regression (21). Thus, compared to (14), data-mining has raised the exponential term because the in-sample fit of the model is improved by eliminating the restriction $\theta = \beta_j$. Moreover, the data-mining procedure replaced the model-specific penalty term $|\tilde{X}_j'\tilde{X}_j/\lambda^2+1|^{-1/2}$ in (14) by $|1/\kappa+1|^{-1/2}$. Thus, provided that

$$\kappa \geq \frac{\lambda^2}{\tilde{X}_j'\tilde{X}_j}, \tag{25}$$

we obtain

$$p(Y|X, X, \tilde{M}_j) \geq p(Y|X, X, M_j). \tag{26}$$

For $\kappa = 1$ condition (25) requires that the prior density in model M_j is more diffuse than the likelihood function, which is a very mild restriction. We conclude that access to the full sample creates an incentive for data-based modifications of the original model M_j because, according to (17),

$$\Pi_{NH}\big(p(Y|X, X, \tilde{M}_j)\big) \geq \Pi_{NH}\big(p(Y|X, X, M_j)\big). \tag{27}$$

The modeler chooses Option (ii) and the reported posterior for θ is $\tilde{p}_j(\theta|Y, X, S) = p(\theta|Y, X, X, \tilde{M}_j)$ given in (23).

Holdout Mechanism. Here the modeler has no information about the holdout sample Y_p. The subjective beliefs of modeler j about $\hat{\theta}_{pm}$ are summarized by the posterior density $p(\hat{\theta}_{pm}|Y_r, X, S, M_j)$. If the modeler chooses Option (i), then his expected payoff is given by

$$\begin{aligned} &\mathbb{E}\left[\Pi_H\big(p(\hat{\theta}_{pm}|Y_r, X, S, M_j)\big)\middle| Y_r, X, S, M_j\right] \\ &= \int \ln[p(\hat{\theta}_{pm}|Y_r, X, S, M_j)]p(\hat{\theta}_{pm}|Y_r, X, S, M_j)d\hat{\theta}_{pm}. \end{aligned} \tag{28}$$

If, one the other hand, the modeler engages in in-sample data mining and chooses Option (ii), then the expected payoff is:

$$\begin{aligned} &\mathbb{E}\left[\Pi_H\big(p(\hat{\theta}_{pm}|Y_r, X, S, \tilde{M}_j)\big)\middle| Y_r, X, S, M_j\right] \\ &= \int \ln[p(\hat{\theta}_{pm}|Y_r, X, S, \tilde{M}_j)]p(\hat{\theta}_{pm}|Y_r, X, S, M_j)d\hat{\theta}_{pm}. \end{aligned} \tag{29}$$

Note that the payoff is a function of the predictive density associated with the data-mined model $\tilde{M}_j$, whereas the expectation is taken using the modeler's subjective beliefs which are based on the original model M_j.

According to a result that dates back at least to Winkler (1969), the compensation scheme based on the log predictive density induces the modeler to reveal his subjective beliefs. To see why, notice that Jensen's inequality implies that

$$\int \left(\ln \left[\frac{p(\hat{\theta}_{pm}|Y_r, X, S, \tilde{M}_j)}{p(\hat{\theta}_{pm}|Y_r, X, S, M_j)} \right] p(\hat{\theta}_{pm}|Y_r, X, S, M_j) \right) d\hat{\theta}_{pm} \leq \ln \left[\int p(\hat{\theta}_{pm}|Y_r, X, S, \tilde{M}_j) d\hat{\theta}_{pm} \right] = 0.$$

Thus, the expected payoff from reporting the results from the Bayesian analysis of model M_j exceeds the expected payoff from data mining and the modeler chooses Option (i) under the holdout mechanism.

So far, we have provided a qualitative characterization of the behavior of the two modelers. The policy maker, in our environment, can now minimize his prediction risk by choosing between the *no-holdout* and the *holdout* mechanism. With regard to the *holdout* mechanism he has to determine the optimal size and composition (in terms of observations from the treatment and control group) of the holdout sample. The next section provides a numerical illustration.

5 Numerical Illustration

This section provides a numerical illustration in which we compare risks under the *holdout* and the *no-holdout* mechanism. For the *holdout* mechanism we consider various sample splitting schemes that differ with respect to the relative sized and composition of the holdout sample. The simulation design is presented in Section 5.1 and the numerical results are discussed in Section 5.2.

5.1 Policy Experiment, Loss Function, and Parameterization

The policy maker is assumed to have conducted an experiment with $n = 1,000$ observations, 500 from a randomly selected treatment group that received the subsidy, $s = \bar{s} = 2$, and 500 are from a control group that did not receive the subsidy, $s = 0$. We reparameterize the prior variance such that $\lambda = \tilde{\lambda}n$. The implication of this reparameterization is that all the

statistics that we compute subsequently have a well-defined limit as the sample size $n \longrightarrow \infty$. Thus, the exact sample size n, used in the simulations is not crucial. Each individual i has two observable characteristics, $x_{i,1}$ and $x_{i,2}$. Let $x_i = [x_{i,1}, x_{i,2}]'$. We assume that

$$x_i \sim iidN(0, \Gamma), \quad \Gamma = \begin{bmatrix} 2 & 0.4 \\ 0.4 & 2 \end{bmatrix}. \tag{30}$$

Thus, the correlation between the two characteristics is 0.2. These assumptions complete the specification of $p(X, S) = p(X)p(S)$.

The policy maker assigns probabilities $\pi_{1,0} = \pi_{2,0} = 1/2$ to the two models M_1 and M_2. We set the precision of the prior densities for θ in models M_1 and M_2 to $\tilde{\lambda}^2 = 1$. This choice of $\tilde{\lambda}$ implies that in a regression of Y on X_j the likelihood function is about twice as informative as the prior.[11] From the policy maker's perspective the distribution of the data takes the form

$$p(Y, X, S) = \frac{1}{2}p(Y, X, S|M_1) + \frac{1}{2}p(Y, X, S|M_2), \tag{31}$$

where

$$p(Y, X, S|M_j) = p(X)p(S) \int p(Y|\theta, X, S, M_j)p(\theta|M_j)d\theta.$$

The unconditional probability, integrating out the data under $p(Y, X, S)$, that the highest posterior probability model corresponds to the "true" model is

$$\begin{aligned} &\pi_{1,0} \int \mathcal{I}_{\{\pi_{1,n} \geq \pi_{2,n}\}} p(Y, X, S|M_1)d(Y, X, S) \\ &+ \quad \pi_{2,0} \int \mathcal{I}_{\{\pi_{1,n} < \pi_{2,n}\}} p(Y, X, S|M_2)d(Y, X, S) = 0.68, \end{aligned} \tag{32}$$

where the posterior model probabilities $\pi_{j,n}$ are functions of (Y, X, S) and $\mathcal{I}_{\{x \geq a\}}$ is the indicator function that is equal to one if $x \geq a$ and is equal to zero otherwise. Thus, as in real-life applications, there is substantial model uncertainty in our simulation design. As mentioned in the introduction, we interpret the fact that the agents are equipped with specifications that cannot be perfectly be distinguished based on the available data as the outcome of the model development stage, that we did not incorporate into our analytical framework. Neither of the two specifications would be strongly rejected by the data.

[11] Omitting the regressor S, the Hessian (multiplied by -1) of the log likelihood function is given by $X_j'X_j \approx 2n$ and the Hessian (multiplied by -1) of the log prior density is $n\tilde{\lambda}^2 = n$.

The policy maker contemplates raising the subsidy from $\bar{s} = 2$, the level in the experiment, to $s^* = 4$. To assess that new policy, the policy maker considers the prediction of the effect of subsidy level s^* on an individual with given characteristics x_1^* and x_2^*. The prediction is evaluated under a quadratic loss function and we will focus on the difference $\Delta(\hat{y})$, defined in (11), between the risk associated with the Bayes prediction and the prediction that the policy maker is able to implement based on the information provided by the modelers.

To make the subsequent exposition more transparent, we take the following short cut. Throughout the analysis we replace model averaging by model selection, restricting the model weights to be zero or one. This leads to the following post-model-selection Bayes predictor

$$\hat{y}^* = \begin{cases} (x_1^* + s^*)\hat{\theta}_1^* & \text{if } \pi_{1,n} \geq \pi_{2,n} \\ (x_2^* + s^*)\hat{\theta}_2^* & \text{otherwise} \end{cases}, \tag{33}$$

where $\hat{\theta}_j^*$ denotes the (full-sample) posterior mean of θ under model M_j, defined in (13). Likewise, the policy maker computes a post-model-selection predictor based on the results elicited from the two modelers:

$$\hat{y} = \begin{cases} (x_1^* + s^*)\hat{\theta}_1 & \text{if } \tilde{\pi}_{1,n} \geq \tilde{\pi}_{2,n} \\ (x_2^* + s^*)\hat{\theta}_2 & \text{otherwise} \end{cases}. \tag{34}$$

$\hat{\theta}_j$ is the posterior mean associated with the reported density $\tilde{p}_j(\theta|Y, X, S)$:

$$\hat{\theta}_j = \begin{cases} \hat{\theta}_j^* & \text{if modeler chooses Option (i)} \\ \tilde{\theta}_j & \text{if modeler chooses Option (ii)} \end{cases},$$

where $\tilde{\theta}_j$ was defined in (23). We set

$$x_1^* = \sqrt{2} \quad \text{and} \quad x_2^* = 0.2\sqrt{2}, \tag{35}$$

i.e., we are setting x_1^* equal to its standard deviation and x_2^* equal to its expected value conditional on $x_1^* = \sqrt{2}$. These choices correspond to the first column of the Cholesky factor of Γ. Note that $x_1^* \neq x_2^*$ generates an automatic penalty for selecting a model that differs from the highest posterior probability model.

The policy maker can choose the size and composition of regression and holdout samples. We characterize the regression sample Y_r in terms of $r \in (0, 1]$, the fraction of the outcome

Table 1: Composition of Estimation Sample Y_r, $n = 1,000$

	$\tau = \tau_{min}$		$\tau = 0.5$	
	Control	Treatment	Control	Treatment
$r = 0.2$	200	0	100	100
$r = 0.5$	500	0	250	250
$r = 0.8$	500	300	400	400
$r = 1.0$	500	500	500	500

data, and τ, the fraction of observations from the treatment group.[12] We restrict our attention to two choices of τ: $\tau = 0.5$ and $\tau = \tau_{min}(r)$, where $nr\tau_{min}(r)$ is the smallest number of observations from the treatment group that can be assigned to the regression sample. If $r = 0.2$ then the regression sample consists of 200 observations (recall $n = 1,000$). Since Y contains 500 observations from the treatment group $\tau_{min}(r) = 0$. If $r = 1$ then $\tau_{min}(r)$ is equal to 0.5. Table 1 summarizes the composition of Y_r for selected values of r and the two choices of τ. By varying r and τ the policy maker can control the variability of the regressor S_r in the regression sample.

In the remainder of this section we analyze the size of integrated risk differentials $\Delta(\hat{y})$ under the *no-holdout* and the *holdout* mechanism. Recall that the integrated risk differential is obtained by averaging over the data (Y, X, S) under the marginal distribution $p(Y, X, S)$ defined in (31). For expositional purposes, we also report results obtained by averaging conditional on a particular value of θ. Under the *holdout* mechanism we can vary the size r of the regression sample and its composition τ. To understand the shape of the risk differential as a function of r and τ we also examine expected frequencies of choosing the highest posterior probability model.

[12]Given the symmetry of our experimental design, it is immaterial whether τ is defined in terms of the treatment or control group.

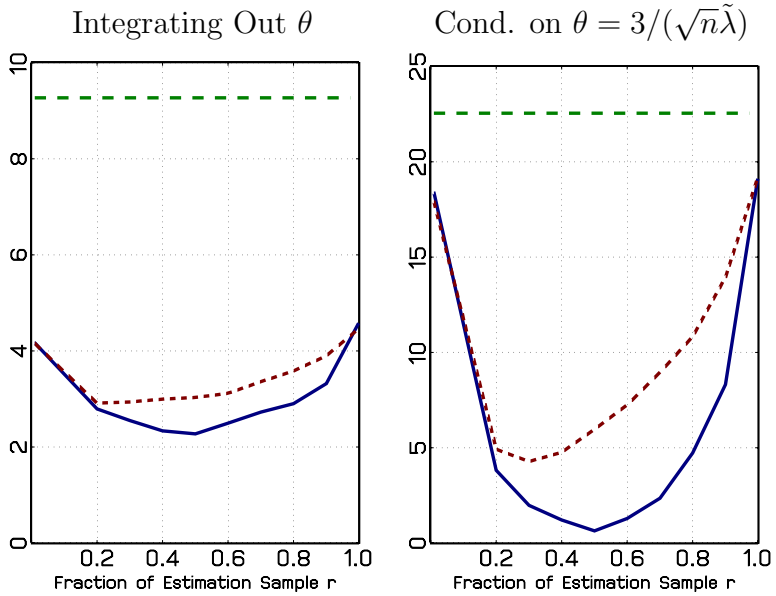
5.2 Ranking of Mechanisms

Figure 1 depicts the risk differentials $\Delta(\hat{y})$, defined in (11), under the *no-holdout* mechanism (long dashed green lines) and under the *holdout* mechanism (solid blue and short dashed red lines, respectively). In the left panel we use the prior distribution to average over all θ values when computing risk differentials, whereas in the right panel we condition on $\theta = 3/(\sqrt{n}\tilde{\lambda})$. This particular value of θ, which is large in the sense that it lies in the far right tail of the prior distribution, amplifies the risk differentials. Two important results emerge from Figure 1. First, the risk differentials are substantially larger under the *no-holdout* mechanism than under the *holdout* mechanism. Second, under the *holdout* mechanism the risk differentials have a U shape as a function of the the size of the estimation sample r. For large values (in absolute terms) of θ, as in the case of the right panel of Figure 1, the U shape is very pronounced whereas the profile is fairly flat for values of θ near zero (not shown in the figure). The differentials are uniformly smaller if the modelers receive the minimum number of observations from the treatment sample. For $\tau = \tau_{min}$ the minimum is obtained for $r = 0.5$, whereas for $\tau = 0.5$ the minimum is achieved at, approximately, $r = 0.3$.

We begin by examining the large risk differentials under the *no-holdout* mechanism. The discrepancy between the predictors $\hat{y}$ and $\hat{y}^*$ can arise from the policy maker not being able to choose the highest posterior probability model based on the information that he receives from the modelers and from discrepancies between the posterior mean estimator $\hat{\theta}_j^*$ and the estimator $\tilde{\theta}_j$ associated with the data-mined model $\tilde{M}_j$. It turns out that the latter discrepancy is the main determinant of the risk differential.

Whenever competing models are *a priori* equally likely, the highest-posterior probability model is the one that attains the highest marginal likelihood $p(Y|X, S, M_j)$, which was given in (14). Because in our simulation the regressors X_1 and X_2 have equal variance and we are using identical priors for θ, the penalty term $|1 + \tilde{X}_j'\tilde{X}_j/\lambda^2|^{-1/2}$ is approximately identical for models M_1 and M_2. Thus, the log marginal likelihood differential is determined by the goodness-of-fit term. Abstracting from the effect of the prior distribution, the goodness-of-fit is determined by the SSR of a (restricted, $\beta_j = \theta$) regression of Y on $\tilde{X}_j = X_j + S$. Under the *no-holdout* mechanism both modelers engage in data mining (Option (ii)) and report

Figure 1: Risk Differentials under *Holdout* and *No-Holdout* Mechanisms



Notes: As risk differentials we report $\Delta(\hat{y})$ defined in (11). In the right panel we replace $p(Y|X,S)$ by $p(Y|X,S,\theta)$ when computing the risk differential. Solid blue and (short) dashed red lines depict outcomes under the *holdout* mechanism, whereas (long) dashed green line depicts outcome under the *no-holdout* mechanism. Fraction of observations from treatment sample: $\tau = \tau_{min}$ is solid blue, $\tau = 0.5$ is (short) dashed red.

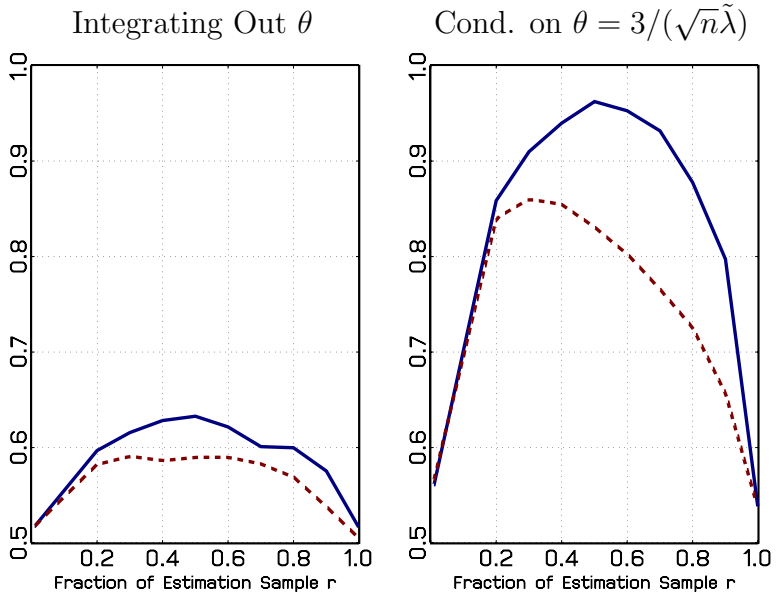
$p(Y|X,S,\tilde{M}_j)$ in (24). The penalty term is identical for the two models and the goodness-of-fit term corresponds to the SSR associated with an (unrestricted) regression of Y on X_j and S. It turns out that in our simulation design the model ranking based on the restricted and unrestricted SSR is almost identical.

The loss differentials under the *no-holdout* mechanism are by and large generated by the discrepancy between $\hat{\theta}_j^*$ and $\tilde{\theta}_j$. Because the assignments to treatment and control groups are independent of the characteristics X_1 and X_2

$$\tilde{\theta}_j \approx (S'S)^{-1}S'Y = \hat{\theta}_{pm}.$$

Thus, the policy maker does not learn anything from engaging the two structural modelers, because the modelers, in essence, abandon the cross-coefficient restrictions implied by their models. In our linear setting, the cost associated with ignoring the cross-coefficient restrictions is increasing in the variability of the regressors X_j relative to the variability of S.

Figure 2: Probability of Finding the Highest-Post.-Prob. Model under *Holdout* Mechanism



Notes: Fraction of observations from treatment sample: $\tau = \tau_{min}$ is solid blue, $\tau = 0.5$ is (short) dashed red.

Given the binary nature of the treatment indicator, the latter is equal to $\bar{s}^2/4 = 1$, whereas the former is set equal to 2 in our simulation design. Overall, this leads to a loss differential of about 9.5 (see Figure 1), which is substantially larger than the loss differential under the *holdout* mechanism.

Under the *holdout* mechanism $\hat{y} = \hat{y}^*$ whenever the policy maker is able to determine the highest-posterior-probability model based on the limited information contained in $p(\hat{\theta}_{pm}|Y_r, X, S, M_j)$. Thus, we will focus on the probability of the policy maker finding the highest-posterior-probability model, which is depicted in Figure 2. As a function of the size r of the regression sample, this probability has an inverted U-shape, which mirrors the U-shaped risk differentials in Figure 1. As before, in the left panel we average over θ when simulating the trajectories (Y, X, S), whereas in the right panel we condition on θ being equal to three prior standard deviations: $\theta = 3/\sqrt{n}\tilde{\lambda}$. Conditional on $\theta = 3/\sqrt{n}\tilde{\lambda}$, the probability that M_j is the highest posterior probability model if data have been generated from M_j is approximately one. Thus, the right panel can also be interpreted as the probability of selecting the "true" model using $p(\hat{\theta}_{pm}|Y_r, X, S, M_j)$.

To understand the inverted U-shape, the following algebraic manipulations are instructive. Given the linear structure of our setup, the distribution of $\hat{\theta}_{pm}|(Y_r, X, S)$ is normal. Because X_1 and X_2 are the same variance and are uncorrelated with S the variance associated with the posterior predictive distribution of $\hat{\theta}_{pm}|(Y_r, X, S)$ is (in large samples) the same for models M_1 and M_2. Thus, differences across models in the predictive distribution are due to differences in the posterior means of $\hat{\theta}_{pm}$. Conditional on M_j the posterior mean of $\hat{\theta}_{pm}$ is given by

$$\mathbb{E}[\hat{\theta}_{pm}|Y, X, S, M_j] = (S'S)^{-1}S' \begin{bmatrix} Y_r \\ \tilde{X}_{j,p}(\tilde{X}'_{j,r}\tilde{X}_{j,r} + \tilde{\lambda}^2)^{-1}\tilde{X}'_{j,r}Y_r \end{bmatrix},$$

where $S = [S'_r, S'_p]'$. This formula highlights that the estimated model is used to predict the missing observations Y_p by

$$\tilde{X}_{j,p}\mathbb{E}[\theta|Y_r, X, S, M_j] = \tilde{X}_{j,p}(\tilde{X}'_{j,r}\tilde{X}_{j,r} + \tilde{\lambda}^2)^{-1}\tilde{X}'_{j,r}Y_r.$$

The imputed observations together with Y_r are then used to predict $\hat{\theta}_{pm}$. Using the definition of $\hat{\theta}_{pm}$, we can write

$$\begin{aligned} &\big(\hat{\theta}_{pm} - \mathbb{E}[\hat{\theta}_{pm}|Y, X, S, M_j]\big)^2 \qquad (36) \\ &= \big(Y'_p - Y'_r\tilde{X}_{j,r}(\tilde{X}'_{j,r}\tilde{X}_{j,r} + \tilde{\lambda}^2)^{-1}\tilde{X}'_{j,p}\big)S_p(S'S)^{-1}S'_p\big(Y_p - \tilde{X}_{j,p}(\tilde{X}'_{j,r}\tilde{X}_{j,r} + \tilde{\lambda}^2)^{-1}\tilde{X}'_{j,r}Y_r\big), \end{aligned}$$

(36) needs to be compared to the goodness-of-fit term that appears in the definition of the marginal likelihood in (14).[13]

Because the goal of the *holdout* mechanism is to discourage data mining, (36) contains no information about the fit of the model on the regression sample Y_r. The forecast errors for Y_p are projected on the space spanned by S_p, which leads to an additional loss of information. If Y_r is small, then the estimate of θ that is used to predict Y_p is very different from the full-sample estimate that underlies the calculation of the goodness-of-fit in (14). In other words, the inverted U-shape in Figure 2 is obtained because for small values of r the selection criterion suffers from imprecise estimates of θ. Large values of r, on the other hand, yield

[13]Recall that we adopted the convention that all regressors have been demeaned. Thus, regardless of r and τ, $S_p \neq 0$.

short holdout samples which make it more difficult to measure the predictive performance of M_1 versus M_2.

The composition of the regression sample, which is controlled by τ, affects the variability of the $\tilde{X}_{j,r}$, $\tilde{X}_{j,p}$, and S_p. If the regression sample is homogeneous, i.e., it only contains observations from the control (or treatment) group, the variance of $\tilde{X}_{j,r}$ is relatively small. For $\tau = \tau_{min}$ the probability of selecting the best model peaks for $r = 0.5$ and almost reaches about 0.65 if θ is integrated out. For $\tau = 0.5$ the probability of selecting the highest-posterior-probability model is less than for $\tau = \tau_{min}$. Conditioning on $\theta = 3/\sqrt{n}\tilde{\lambda}$, the effect of the composition of the regression sample is more pronounced. For $\tau = \tau_{min}$ the probability peaks at $r = 0.5$ and almost reaches one. For $\tau = 0.5$ the probability peaks at $r = 0.3$ but only reaches about 0.85. In sum, we conclude that in our environment the *holdout* mechanism dominates the *no-holdout* mechanism and it is best for the policy maker to set $r = 0.5$ and $\tau = \tau_{min}$, that is, to provide the modeler with the control (or treatment) sample only.

The result about the optimal composition of the holdout sample may appear counter-intuitive. Mixing observations from the control and the treatment group in the holdout sample increases the precision of the estimate of θ. However, to mimic the full Bayesian solution it is not the precision of the treatment effect *per se* that matters. Instead, the model ranking based on $\left(\hat{\theta}_{pm} - \mathbb{E}[\hat{\theta}_{pm}|Y, X, S, M_j]\right)^2$ has to match the model ranking based on the goodness-of-fit term in (14). Combining control and treatment group observations increases the correlation among the regressors $\tilde{X}_{1,r}$ and $\tilde{X}_{2,r}$ (recall that $\tilde{X}_j = X_j + S$) which makes it more difficult to distinguish the model specifications. To assess the robustness of this finding, we reduced the variance of the regressors X_1 and X_2 by a factor of 10 in the simulation design, which makes it more difficult to identify θ based on a homogeneous regression sample. Nonetheless, the result was qualitatively the same: the choice of $\tau = \tau_{min}$ dominates $\tau = 0.5$.

The exact magnitude of the loss differentials is, of course, sensitive to the parameterization of the experiment. Moreover, in actual applications the structural models are much more complicated than the simple linear regressions considered in this paper. Nonetheless,

we believe that our setup captures the essence of the problem. The relative variability of X_j and S captures the information content in the cross-equation restrictions, which in practice could be large or small. In our illustration, the posterior variance of θ under the restricted ($\beta_j = \theta$) regression is twice as large as under the unrestricted regression. The correlation between X_1 and X_2 controls the similarity of the two structural models. The higher the correlation, the more similar the models, and the lower the stakes in determining the highest-posterior-probability model, because both models will deliver very similar predictions. If both modelers engage in data mining, then the estimates of the treatment effect given the observed treatment level $\bar{s}$ are identical across modelers and identical to the simple difference-in-means estimator that the policy maker is able to compute himself. In this sense, data mining makes the two structural models more similar. In more realistic settings the overall fit of the two data-mined models would probably also become more similar. In our stylized framework the overall fit remains different because we impose that the two modelers use different regressors.

6 Conclusion

We developed a principal-agent framework that allows us to characterize potential costs of data mining and potential benefits of holdout samples designed to discourage data mining. In our environment the full Bayesian posterior mean prediction is first-best. However, the tasks of decision making and model estimation is divided among a policy maker and a set of modelers. The policy maker would like to implement the first-best Bayesian decision. To that end, it is assumed that the modelers are rewarded based on the fit of the models that they provide. This compensation scheme creates an incentive for the modelers to engage in data-mining and to overstate the fit of their models. In our numerical illustration we find that the policy maker minimizes risk by withholding 50% of the sample from the modelers and only makes available observations either from the control group or the treatment group.

Holdout samples have not, to our knowledge, been used by actual policy makers as a tool for model selection. Indeed, in the few examples based on randomized controlled trials

(RCT), the use of a holdout sample has been initiated by the researchers themselves.[14] In those cases, having access to data from both the treatment and control groups, researchers have chosen holdout samples comprised of observations solely from one or the other group rather than observations from a mixture of both groups. This choice is consistent with the findings from our numerical illustration. For our results to apply, however, it must be assumed that those researchers acted exactly as the modelers in our setting, that is, as if they did not have access to the holdout sample during the estimation of their models.

Although our results are based on a numerical illustration, it is our speculation that they would hold more generally, at least in the RCT setting. If that is the case, then we would also argue that the use of a holdout sample given data from an RCT (a growing empirical methodology) should be standard practice. We believe that if this practice were established profession-wide, researchers would maintain the necessary distinction between the estimation sample and the holdout sample.

References

Attanasio, O., C. Meghir, and A. Santiago (2012): "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA," *Review of Economic Studies*, 79, 37–66.

Bernardo, J., and A. Smith (1994): *Bayesian Theory*. John Wiley & Sons New York.

Clemen, R., and R. Winkler (2007): "Aggregating Probability Distributions," in *Advances in Decision Analysis: From Foundations to Applications*, ed. by W. Edwards, R. Miles, and D. von Winterfeldt, pp. 154–176. Cambridge University Press.

Duflo, E., R. Hanna, and S. Ryan (2011): "Incentives Work: Getting Teachers to Come to School," *American Economic Review*, forthcoming.

Ferrall, C. (2012): "Explaining and Forecasting Results of the Self-sufficiency Project," *Review of Economic Studies*, 79, 1495–1526.

[14] In some cases, the RCT is itself conducted by the researcher.

French, S. (1985): "Group Consensus Probability Distributions: A Critical Survey," in *Bayesian Statistics 2*, ed. by J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, pp. 183–202. Elsevier.

Genest, C., and J. Zidek (1986): "Combining Probability Distributions: A Critique and Annotated Bibliography," *Statistical Science*, 1(1), 114–135.

Ichimura, H., and C. Taber (2000): "Direct Estimation of Policy Impacts," *NBER Technical Working Paper*, 254.

Lamont, O. (2002): "Macroeconomic Forecasts and Microeconomic Forecasters," *Journal of Economic Behavior & Organization*, 48, 265–280.

Laster, D., P. Bennet, and I. S. Geoum (1999): "Rational Bias in Macroeconomic Forecasts?," *Quarterly Journal of Economics*, 114, 293–318.

Leamer, E. (1978): *Specification Searches – Ad Hoc Inference with Nonexperimental Data*. John Wiley & Sons New York.

Lindley, D. V. (1985): "Reconciliation of Discrete Probability Distributions," in *Bayesian Statistics 2*, ed. by J. Bernardo, M. DeGroot, D. Lindley, and A. Smith, pp. 375–390. Elsevier.

Lo, A., and C. MacKinlay (1990): "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies*, 3(3), 431.467.

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

Sandroni, A. (2003): "The Reproducible Properties of Correct Forecasts," *International Journal of Game Theory*, 32, 151–159.

Schorfheide, F., and K. Wolpin (2012): "On the Use of Holdout Samples for Model Selection," *American Economic Review: Papers and Proceedings*, 102(3), 477–481.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6(2), 461–464.

Stone, M. (1977): "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion," *Journal of the Royal Statistical Society Series B*, 39(1), 44–47.

Todd, P., and K. Wolpin (2006): "Assessing the Impact of a Child Subsidy Program in Mexico: Using a Social Experiment to Valdidate a Behavioral Model of Child Schooling and Fertility," *American Economic Review*, 96(5), 1384–1417.

——— (2008): "Ex Ante Evaluation of Social Programs," *Annales d'Economie et de Statistique*, 91-92, 263–292.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097–1126.

Winkler, R. (1969): "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association*, 64(327), 1073–1078.

Wise, D. (1985): "Behavioral Model versus Experimentation: The Effects of Housing Subsidies on Rent," in *Methods of Operations Research 50*, ed. by P. Brucker, and R. Pauly, pp. 441–489. Königstein: Verlag Anton Hain.

Wolpin, K. (2013): *The Limits of Inference without Theory*. MIT Press.