To Hold Out or Not to Hold Out

Frank Schorfheide* Kenneth I. Wolpin

University of Pennsylvania *University of Pennsylvania*

December 14, 2011

Abstract

Researchers often hold out data from the estimation of econometric models to use for external validation. However, the use of holdout samples is suboptimal from a Bayesian perspective, which prescribes using the entire sample to form posterior model probabilities and predictive distributions for policy effects. In this paper, we develop a principal-agent framework, in which a first-best Bayesian solution is not implementable because econometric modelers engage in data-mining and misrepresent the fit of their models. A policy-maker can use a holdout sample to discourage data mining. Using a stylized representation of a randomized controlled trial, we set up a simulation experiment to examine under what conditions model weighting based on predictions for a holdout sample could be justified, how large it should be and how it should mix observations from treatment and control groups. (JEL C11, C31, C52)

KEY WORDS: Bayesian Analysis, Model Selection, Principle-Agent Models, Randomized Controlled Trials.

*Correspondence: University of Pennsylvania, Department of Economics, 3718 Locust Walk, Philadelphia, PA 19104-6297. Email: schorf@ssc.upenn.edu (F. Schorfheide); wolpink@ssc.upenn.edu (K. Wolpin). We thank George Mailath as well as seminar participants at Princeton and the University of Pennsylvania for helpful comments and suggestions.

1 Introduction

The use of Randomized controlled trials (RCTs) to evaluate policies has become a prominent methodology in applied economics. An important limitation is that one cannot extrapolate outside of the treatment variation in the particular experiment. Moreover, given their cost, RCTs cannot be used to perform *ex ante* policy evaluation over a wide range of policy alternatives. Thus, extrapolation to new treatments requires developing and estimating models that embed behavioral and statistical assumptions, what we will refer to as structural models. It is therefore important to have methods for assessing the relative credibility of competing structural models.

In practice researchers often hold out data from estimation to use for external validation, e.g., Wise (1985), Todd and Wolpin (2006), and Duflo, Hanna, and Ryan (2011) in the case of RCTs. Although having intuitive appeal, the use of holdout samples is puzzling from a Bayesian perspective, which prescribes using the entire sample to form posteriors. The contributions of this paper are twofold. First, we provide a formal, albeit stylized, framework in which data mining poses an impediment to the implementation of the ideal Bayesian analysis. Data mining, for the purpose of this paper, is a process by which a modeler tries to improve the fit of a structural model during estimation, e.g., change functional forms, add latent state variables. 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 ideal Bayesian solution.

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. We assume that the modelers get rewarded in terms of the fit of their model. Two mechanisms are considered. Under the *no-holdout* mechanism the modelers get 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. Since the modelers have access to the full sample, there exists 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 hold-out mechanism is designed so that the modelers truthfully reveal their subjective beliefs about the hold-out sample. However, predictive distributions for the hold-out 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.

While 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 and that the lowest level of risk is attained by holding back 50% of the sample and providing the modelers only with data either from the control or from the treatment group.

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 pay-off 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 pay-off 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).

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 statis-

tical 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. relaxing function form restrictions, 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. A review of the Bayesian analysis of the linear Gaussian regression model as well as proofs and derivations are provided in the Appendix.

2 A Working Example

In order 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.[1] It is assumed that prior to the policy change there is no direct tuition cost of schooling. The goal is to determine an optimal level of the tuition 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, $s = 0$. Suppose that the outcome variable for household i, $i = 1, \ldots, n$, is denoted by y_i and is continuous, e.g. attendance measured in hours. In addition to the outcome, two scalar characteristics $x_{i,1}$ and $x_{i,2}$ as well as the level of treatment $s_i \in \{0, \bar{s}\}$ are observed for each household. Let $x_i = [x_{i,1}, x_{i,2}]$.

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

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

Here $U_j(\cdot)$ is a model-specific utility function, parameterized in terms of ϑ_j, c is consumption, $h \in [0, 1]$ is hours spent in school (the total endowment of time has been normalized to one), z is a vector of observable household characteristics, ϵ is a random variable that captures unobservable heterogeneity, and inc is parental income.

We denote the optimal attendance decision by $h = \varphi_j(inc, w; z, \epsilon, \vartheta_j)$. An attendance

[1] 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).

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

subsidy s modifies the households' budget constraint to

$$c = inc + w(1-h) + sh = (inc + s) + (w - s)(1 - h) = \widetilde{inc} + \tilde{w}(1-h). \tag{2}$$

The optimal attendance choice in the presence of a subsidy is

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

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. 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:

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

The j subscripts in (4) capture the different assumption embodied in the two models about the relevant characteristic x that affects the outcome. The error term u_i arises from the heterogeneity generated by the unobserved characteristics ϵ in (1). 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 $\theta = \beta_j$.[3] Thus, variation in $x_{i,j}$ is sufficient to obtain a measurement of the subsidy effect. Since we will subsequently use matrix notation, let X_j be 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.

3 Bayesian Analysis

Throughout this paper we adopt a Bayesian approach to analyze the problem of determining an optimal subsidy level, which requires us to specify priors for the parameters of models

[3] 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 $\theta = -\beta_j$.

M_1 and M_2 as well as to the models themselves. Both models are equipped with the prior distribution $\theta \sim N\big(0, 1/(n\lambda^2)\big)$. The density of this prior is denoted by $p(\theta|M_j)$. Overall, this leads to

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

where $\tilde{X}_j = X_j + S$. We use the scaling of the prior variance by $1/n$ as a technical device to ensure that the models do not become perfectly distinguishable as the sample size tends to infinity (see below for further discussion). We assume that the marginal density of (X, S) does not depend on θ and is the same for both models. As a consequence, $p(X, S)$ cancels from most of the formulas presented below and results are presented in terms of densities that are conditional on (X, S). Given randomization, the selection of the treatment group is independent of the observable characteristics, that is, $p(X, S) = p(X)p(S)$. 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

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

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 y_* for an individual that receives a subsidy s_* and has characteristics $x_1 = x_2 = x_*$ under a quadratic loss. The integrated risk associated with the prediction $\hat{y}$ is

$$\mathcal{R}(\hat{y}) = \int_{Y,X,S} \left[\int_\theta (\theta(x_* + s_*) - \hat{y})^2 p(\theta|Y, X, S) d\theta \right] p(Y|X, S) p(X, S) d(Y, X, S). \tag{7}$$

The integrated risk is minimized by minimizing the posterior expected loss (the term in square brackets in (7)) for each sample (Y, X, S). This leads to the posterior mean predictor

$$\hat{y}_* = \int \theta(x_* + s_*) \left(\sum_{j=1,2} \pi_{j,n} p(\theta|Y, X, S, M_j) \right) d\theta. \tag{8}$$

This Bayesian solution is first-best in our environment and serves as a benchmark in the subsequent analysis.

To calculate the optimal predictor in (8) 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

$$p(\theta|Y,X,S,M_j) = \frac{p(Y|X,S,\theta,M_j)p(\theta|M_j)}{p(Y|X,S,M_j)}. \tag{9}$$

The model specification in (5) implies that this posterior distribution takes the form

$$\theta|(Y,X,S,M_j) \sim N\Big((\tilde{X}_j'\tilde{X}_j+\lambda^2)^{-1}\tilde{X}_j'Y, (n\lambda^2+\tilde{X}_j'\tilde{X}_j)^{-1}\Big). \tag{10}$$

The posterior model probabilities $\pi_{j,n}$ are a function of the marginal likelihoods

$$p(Y|X,S,M_j) = \int_{\theta\in\Theta} p(Y|\theta,X,S,M_j)p(\theta|M_j)d\theta. \tag{11}$$

For linear Gaussian regressions the marginal likelihoods can be calculated analytically and take the form

$$\begin{aligned} p(Y|X,S,M_j) &= (2\pi)^{-n/2}|1+\tilde{X}_j'\tilde{X}_j/(n\lambda^2)|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[Y'(I-\tilde{X}_j(\tilde{X}_j'\tilde{X}_j+n\lambda^2)^{-1}\tilde{X}_j')Y]\right\}. \end{aligned} \tag{12}$$

The exponential term captures the goodness of in-sample fit, whereas the term $|1+\tilde{X}_j'\tilde{X}_j/(n\lambda^2)|^{-1/2}$ 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 θ and thus the prior is in relative terms more diffuse.

4 A Principal-Agent Problem

The policy analysis described in the previous section involves two stages. In the first stage, the two models are estimated and posterior distributions are computed. This leads to the model-conditional posteriors $p(\theta|Y,X,S,M_j)$ in (10) and marginal likelihoods $p(Y|X,S,M_j)$

in (12). In the second stage, the output from the two models is combined via Bayesian model averaging, see (6), and the model mixture is used to generate a policy prediction, see (8). In practice, different individuals might be involved into the two stages of the analysis which potentially creates incentive problems. These incentive problems, in turn, can provide a rationale for holdout samples.

In the remainder of this paper, it is assumed that the first stage of the analysis is conducted by two modelers (agents) and the second stage is executed by a policy maker (principal). 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.

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 three options that are available to the modelers: (i) Bayesian analysis of model M_j based on the data provided by the policy maker; (ii) in-sample data mining, which is represented by a stylized modification of the prior distribution; or (iii) the analysis of a mixture of models that includes a more flexible reference model. Finally, we discuss how the mechanisms affect the modelers' choices 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 is unable to conduct an analysis of the structural models M_1 and M_2. He can only estimate the treatment effect in the experiment 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

$$M_p: \quad Y = S\theta + V, \tag{13}$$

where V is a $n \times 1$ vector of error terms. The resulting estimator of the treatment effect is

$$\hat{\theta}_p = (S'S)^{-1}S'Y. \tag{14}$$

We assume, however, that the policy maker's statistical model M_p cannot be used to extrapolate the treatment effect to other levels of treatment $s \neq \bar{s}$.

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 (7). Since the loss function is quadratic, the integrated risk can be expressed as

$$\mathcal{R}(\hat{y}_p) = \mathcal{R}(\hat{y}_*) + \Delta(\hat{y}_*, \hat{y}_p), \tag{15}$$

where the discrepancy function is given by

$$\Delta(\hat{y}_*, \hat{y}_p) = \int \left(\hat{y}_* - \hat{y}_p\right)^2 p(Y|X,S)p(X,S)d(Y,X,S). \tag{16}$$

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 (10) and (12) can the policy maker implement the first-best Bayesian decision. The compensation of the modelers is a function of how well their models are able to fit the data, adjusting for model complexity. We assume that the compensation is proportional to the reported log marginal likelihood $\ln \tilde{p}_j(Y_p|Y_r, X, S)$. The policy maker updates the model weights according to

$$\tilde{\pi}_{j,n} = \frac{\pi_{j,0}\tilde{p}_j(Y|X,S)}{\pi_{1,0}\tilde{p}_1(Y|X,S) + \pi_{2,0}\tilde{p}_2(Y|X,S)}. \tag{17}$$

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 modelers for estimation purposes and Y_p is a *holdout* or *prediction* sample that can be used by the policy maker to evaluate predictions.[4] The mechanism unfolds in two stages. First, the policy maker asks the modelers to provide a predictive density $\tilde{p}_j(\hat{\theta}_p|Y_r, X, S)$ for his estimate of the treatment effect given by (14). This predictive density is then used to update the model probabilities. 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 $\tilde{p}_j(\theta|Y, X, S)$. We assume that the compensation is proportional to $\ln \tilde{p}_j(\hat{\theta}_p|Y_r, X, S)$. The policy maker updates the model weights according to

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

A few remarks about the assumptions on the mechanisms are in order. First, the pay-off for Modeler 1 is independent of the action taken by Modeler 2, and vice versa. Thus, we abstract from strategic interactions between the modelers. Second, for the *holdout* mechanism we assumed that the policy maker updates the model weights based on the predictive densities $\tilde{p}_j(\hat{\theta}_p|Y_r, X, S)$ for the reduced-form 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

[4]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.

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 reduced-form estimate $\hat{\theta}_p$, 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 weights based on the $\hat{\theta}$-predictive density, as for example in Todd and Wolpin (2006) and Duflo, Hanna, and Ryan (2011).

Finally, for the *holdout* mechanism, we assume that the policy maker gives the modelers access to the entire sample once he has determined the model weights. 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 weights due to data-mining. We use $\tilde{p}_j(\theta|Y, X, S)$ to denote the posterior of θ reported by the modelers.

4.3 The Choice Set of the Modelers

We assume that the modelers can choose between the following three options: (i) report results from the Bayesian analysis of M_j based on the sample provided by the policy maker; (ii) introduce a reference model M_{j0} to account for the possibility of model misspecification and report the fit from the mixture of M_j and M_{j0}; (iii) engage in data-mining to improve the fit of model M_j.

Option 1: Report Results from 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 (12). 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 $\hat{\theta}_p(Y_p, Y_r)$, denoted by $p(\hat{\theta}_p|Y_r, X, S, M_j)$. The corresponding full-sample posterior for θ is given by by $p(\theta|Y, X, S, M_j)$.

Option 2: Bayesian Inference with Reference Model. Suppose that the modelers entertain the possibility that their models M_j are misspecified. While there are several

options for introducing concern about misspecification in a Bayesian framework, we assume that the modelers consider a reference model M_{j0} that takes the form of an unrestricted regression with regressors X_j and S

$$M_{j0}: \quad Y = \beta_j X_j + \theta S + U, \quad \beta_j \sim N(0, n\lambda^2), \quad \theta \sim N(0, n\lambda^2). \tag{19}$$

Taking the reference model into account, the modeler constructs the model mixture $\bar{M}_j$:

$$\bar{M}_j: \quad \bar{\pi}_{j,0} p(Y, \theta | X, S, M_j) + \bar{\pi}_{j0,0} p(Y, \theta | X, S, M_{j0}), \tag{20}$$

from which one can compute either the marginal likelihood function $p(Y|X, S, \bar{M}_j)$ or the predictive density $p(\hat{\theta}_p|Y_r, X, S, \bar{M}_j)$. The corresponding full-sample posterior for θ is given by

$$p(\theta|Y, X, S, \bar{M}_j) = \bar{\pi}_{j,n} p(\theta|Y, X, S, M_j) + \bar{\pi}_{j0,n} p(\theta|Y, X, S, M_{j0}), \tag{21}$$

where $\bar{\pi}_{j,n}$ and $\bar{\pi}_{j0,n}$ are posterior probabilities for the original model M_j and the reference model M_{j0}.

Option 3: In-Sample Data-Mining. We represent in-sample data mining as data-based modification of the prior distribution associated with model M_j. This modification breaks the tight link between θ and β and 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 to include in the model and which to leave out, without accounting for this specification search subsequently.

In our working example, the data-mining prior is constructed as follows. We begin by breaking the link between θ and β by introducing an additional parameter ψ such that

$$\beta_j = \theta + \psi_j.$$

A value $\tilde{\psi}_j$ is chosen such that at the posterior mean $\hat{\beta}_j = \hat{\theta} + \tilde{\psi}_j$. Specifically, using the generalized relationship between β_j and θ, the decision rule in (4) leads to the regression

$$Y = X_j(\theta + \psi_j) + S\theta + U = \tilde{X}_j \theta + X_j \psi_j + U, \tag{22}$$

where $\tilde{X}_j = X_j + S$. Define the matrix $M_{\tilde{X}_j} = I - \tilde{X}_j(\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_j'$ and let $\tilde{\psi}_j$ the least squares estimate of ψ_j in (22):

$$\tilde{\psi}_j = (X_j' M_{\tilde{X}_j} X_j)^{-1} X_j' M_{\tilde{X}_j} Y.$$

Plugging $\tilde{\psi}_j$ into (22) yields the modified regression

$$Y = \tilde{X}_j\theta + X_j\tilde{\psi}_j + U. \tag{23}$$

After having relaxed the restriction $\theta = \beta_j$ in a data-driven manner, the prior for θ is centered at the maximum likelihood estimate derived from (23). Its covariance matrix is chosen to be proportional to $(\tilde{X}_j'\tilde{X}_j)^{-1}$. Using the definition $\tilde{Y}_j = Y - X_j\tilde{\psi}_j$ the data-mined model takes the form

$$\begin{aligned} \tilde{M}_j \quad : \quad & Y = \tilde{X}_j\theta + X_j\tilde{\psi}_j + U, \quad \theta \sim N\left(\tilde{\theta}_i, (\kappa\tilde{X}_j'\tilde{X}_j)^{-1}\right) \\ & \tilde{\theta}_j = (\tilde{X}_j'\tilde{X}_j)^{-1}\tilde{X}_i'\tilde{Y}_j. \end{aligned} \tag{24}$$

The parameter κ scales the prior precision of θ. 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 $p(\hat{\theta}_p|Y_r, X, S, \tilde{M}_j)$. 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 (10). By construction, the posterior is centered at $\tilde{\theta}_j$ and takes the form:

$$\theta|Y, X, S, \tilde{M}_j \sim N\left(\tilde{\theta}_j, \left((\kappa+1)\tilde{X}_j'\tilde{X}_j\right)^{-1}\right). \tag{25}$$

While models $\tilde{M}_j$ and M_{j0} are very similar in that they both relax the restriction $\beta = \theta$ in order to improve fit, under *in-sample data-mining* 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. The modelers could report $p(Y|X,S,M_j)$, $p(Y|X,S,\tilde{M}_j)$, or $p(Y|X,S,\bar{M}_j)$. The marginal likelihood associated with M_j is given in (12). The marginal likelihood function of the data-mined model $\tilde{M}_j$ takes the form

$$\begin{aligned} p(Y|X,S,\tilde{M}_j) &= (2\pi)^{-n/2}|1/\kappa+1|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[\tilde{Y}_j'(I-\tilde{X}_j(\tilde{X}_i'\tilde{X}_j)^{-1}\tilde{X}_j')\tilde{Y}_j]\right\}. \end{aligned} \tag{26}$$

Our particular choice of $\tilde{\psi}_j$ ensured that $\tilde{\theta}_j$ corresponds to the Gaussian maximum likelihood estimator of θ in the unrestricted regression (22). Thus, compared to (12) data-mining step 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 $|\tilde{X}_j'\tilde{X}_j/(n\lambda^2)+1|^{-1/2}$ in (12) by $|1/\kappa+1|^{-1/2}$. Thus, provided that

$$\kappa \geq \frac{n\lambda^2}{\tilde{X}_j'\tilde{X}_j}, \tag{27}$$

we obtain

$$p(Y|X,X,\tilde{M}_j) \geq p(Y|X,X,M_j). \tag{28}$$

Thus, data-mining unambiguously raises the marginal likelihood compared to the original model M_j. For $\kappa = 1$ condition (27) requires that the prior density in model M_j is more diffuse than the likelihood function, which is a very mild restriction.

The marginal likelihood function of the model mixture $\bar{M}_j$ is a convex combination of $p(Y|X,S,M_j)$ and $p(Y|X,S,M_{j0})$. If one defines $\bar{X}_j = [X_j, S]$, then the marginal likelihood of the reference model can be expressed as

$$\begin{aligned} p(Y|X,S,M_{j0}) &= (2\pi)^{-n/2}|I+\bar{X}_j'\bar{X}_j/(n\lambda^2)|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[Y'(I-\bar{X}_j(\bar{X}_j'\bar{X}_j+n\lambda^2 I)^{-1}\bar{X}_j')Y]\right\}. \end{aligned} \tag{29}$$

The goodness-of-fit term in (29) corresponds to a regression of Y on X_j and S. But due to the influence of the prior distribution $p(\theta,\beta|M_{j0})$, it is smaller than the goodness-of-fit term for the data-mined model in (26). Thus, under some mild restrictions on λ we can deduce that $p(Y|X,X,\tilde{M}_j) \geq p(Y|X,X,M_{j0})$ and therefore

$$p(Y|X,X,\tilde{M}_j) \geq p(Y|X,X,\bar{M}_j) \tag{30}$$

To summarize, among the three available options, the modelers maximize their pay-off through in-sample data mining and setting $\tilde{p}_j(Y|X,S) = p(Y|X,S,\tilde{M}_j)$. Access to the full sample creates an incentive for data-based modifications of the original model M_j. The reported posterior for θ is $\tilde{p}_j(\theta|Y,X,S) = p(\theta|Y,X,X,\tilde{M}_j)$ given in (25).

Holdout Mechanism. Here the modeler has no information about the holdout sample Y_p. Suppose that the subjective beliefs of the modelers are described by the specification $\bar{M}_j$ which allows for the possibility that the restriction $\theta = \beta_j$ is potentially misspecified. The expected pay-off of the modeler is given by

$$\int \ln[\tilde{p}_j(\hat{\theta}_p|Y_r,X,S)]p(\hat{\theta}_p|Y_r,X,S,\bar{M}_j)d\hat{\theta}_p. \tag{31}$$

According to Jensen's inequality

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

Recall that we introduced the notation $\tilde{p}_j(\cdot)$ to denote the predictive density that is reported to the policy maker and $p(\hat{\theta}_p|Y_r,X,S,\bar{M}_j)$ refers to the predictive density under model $\bar{M}_j$. The inequality turns into an equality if $\tilde{p}_j(\hat{\theta}_p|Y_r,X,S) = p(\hat{\theta}_p|Y_r,X,S,\bar{M}_j)$. Thus, we exploited the fact that a compensation scheme based on the log predictive density induces the econometrician to reveal his subjective beliefs. This result dates back at least to Winkler (1969) and has been credited to unpublished work by Bruno DeFinetti and Leonard Savage. Note, however, that the policy maker would want the modeler to reveal results from M_j instead of $\bar{M}_j$.

The predictive distribution associated with the model mixture $\bar{M}_j$ takes the following form

$$p(\hat{\theta}_p|Y_r,X,S,\bar{M}_j) = \bar{\pi}_{j0,r}p(\hat{\theta}_p|Y_r,X,S,M_{j0}) + \bar{\pi}_{j,r}p(\hat{\theta}_p|Y_r,X,S,M_j), \tag{32}$$

where $\bar{\pi}_{j0,r}$ and $\bar{\pi}_{j,r}$ are posterior weights based on Y_r. If the modelers are provided with a subsample Y_r that contains data from both the treatment and the control group, they can potentially assess their restrictions $\theta = \beta_j$. If the regression sample Y_r provides strong evidence against the restriction $\beta_j = \theta$, then the posterior probability of M_{j0} is close to one and the mixture is dominated by the reference model. Conversely, if the data provide

strong evidence in favor of the more parsimonious model M_j the mixture will resemble $p(\hat{\theta}_p|Y_r, X, S, M_j)$. We assume that the prior probability $\bar{\pi}_j$ is larger than 1/2, such that if Y_r contains no information from the treatment sample, then modelers have no evidence against $\theta = \beta_j$ and reveal M_j. To summarize, under the *holdout* mechanism the modelers report $\tilde{p}_j(\hat{\theta}_p|Y_r, X, S) = p(\hat{\theta}_p|Y_r, X, S, \bar{M}_j)$ and $\tilde{p}_j(\theta|Y, X, S) = p(\theta|Y, X, S, \bar{M}_j)$.

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 loss 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 analyze the modeler's choices as well as the policy maker's loss for various sample splitting choices. The simulation design is presented in Section 5.1. Section 5.2 proceeds by examining the behavior of the modelers under the *holdout* mechanism. In particular, we study the weight that the reference model M_{j0} receives in the predictions of the modelers. Finally, the main results about the policy maker's risk under the *no-holdout* and the *holdout* mechanism for various choices of holdout samples are presented in Section 5.3.

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, $s = 0$. The exact sample size is not important because by making the prior variances proportional to $1/n$ all the statistics

that we subsequently compute have well-defined limits as $n \longrightarrow \infty$. 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

$$x_i \sim iidN(0, \Gamma), \quad \Gamma = \begin{bmatrix} 2 & 0.4 \\ 0.4 & 2 \end{bmatrix}. \tag{33}$$

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. From the policy maker's perspective the distribution of the data takes the form

$$p(Y, X, S) = \frac{1}{2}p(Y, X, S|M_1) + \frac{1}{2}p(Y, X, S|M_2), \tag{34}$$

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 policy maker contemplates raising the subsidy from $\bar{s} = 2$, the level in the experiment, to $s_* = 4$. To assess that 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 evaluate the expected discrepancy (16) between the Bayes prediction and the prediction that the policy maker is able to implement based on the information provided by the modelers. Here the expectation is taken with respect to the marginal density of the observations (Y, X, S).

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

$$\hat{y}_* = \begin{cases} \hat{\theta}(M_1)(x_1 + s_*) & \text{if } \pi_{1,n} \geq \pi_{2,n} \\ \hat{\theta}(M_2)(x_2 + s_*) & \text{otherwise} \end{cases}, \tag{35}$$

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

$$\hat{y}_* = \begin{cases} \tilde{\theta}_1(x_1 + s_*) & \text{if } \tilde{\pi}_{1,n} \geq \tilde{\pi}_{2,n} \\ \tilde{\theta}_2(x_2 + s_*) & \text{otherwise} \end{cases}, \tag{36}$$

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

To complete the specification of the numerical illustration we set the precision of the prior densities for θ in models M_1 and M_2 to $\lambda^2 = 1$. Moreover, we assume that the modelers assign prior probability $\bar{\pi}_{j0} = 0.48$ to the reference model M_{j0}. We make the simplifying assumption that the modelers, like the policy maker, do not average predictions from M_j and M_{j0}. Instead, under the *holdout* mechanism they select 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 data, and τ, the fraction of observations from the treatment group.[5] We restrict our attention to two choices of τ: $\tau = 0.5$ and $\tau = \tau_{min}(r)$, where $nr\tau_{min}$ 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.

In the remainder of this section we analyze expected frequencies of model choices and risk differentials (16) with respect to r and τ. Expected losses are computed under the marginal distribution of (Y, X, S) given by (34) as well as under the conditional distribution (Y, X, S) given that M_1 is "true" and θ is equal to a multiple of the prior standard deviation: $\theta = 0.2/\sqrt{n\lambda^2}$ or $\theta = 5/\sqrt{n\lambda^2}$, where $n\lambda^2 = 1,000$. Given our assumptions about the covariance matrix Γ of the individual characteristics under the first choice of θ, models M_1 and M_2 are fairly difficult to distinguish, whereas the second choice of θ leads to decisive posterior model probabilities. In fact, conditional on $\theta = 5/\sqrt{n\lambda^2}$ the probability that the highest posterior probability model and "true" model coincide is nearly 1.0, because for large values of θ the misspecification associated with the restriction $\beta_2 = \theta$ in model M_2 is easily detectable. The unconditional probability, averaging over θ with respect to the prior

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

distribution, of the highest posterior probability model being the "true" model is 0.68.

5.2 Behavior of Modelers under *Holdout* Mechanism

Under the *holdout* mechanism the modelers only have access to a fraction of the outcome data, they determine based on the posterior odds of M_j versus M_{j0} whether to report results from one or the other model. In order to obtain the correct posterior prediction of the treatment effect, the policy maker would want the modelers to always report results from M_j. Thus, for brevity, we shall refer to a modeler who does indeed report results from M_j as "honest." Suppose that M_1 is the highest posterior probability model. The following outcomes are possible:

(i) Modeler 1 is "honest" and M_1 is selected. In this case $\hat{y}_p = \hat{y}_*$ and the policy maker is able to recover the Bayes predictor of the treatment effect.

(ii) Modeler 1 is not "honest" and policy maker ends up selecting M_{10} because the fit of M_{10} dominates the fit of the model reported by the second modeler. In this case the policy maker's prediction is based on the correct covariate $x_{i,1}$, but it deviates from $\hat{y}_*$ because the restriction $\theta = \beta_j$ is not imposed.

(iii) Modeler 2 is not "honest" and policy maker ends up selecting M_{20} because the marginal likelihood of M_{20} exceeds the marginal likelihood of M_1. This leads to an inferior prediction because the policy maker uses the wrong covariate, x_2 instead of x_1, and misses the restriction between treatment effect and the coefficient on the covariate.

We now study the probabilities that modelers M_1 and M_2 are "honest." Initially, these probabilities are computed conditional on model M_1 being "true" and a particular value of θ. As explained above, we consider θ being equal to 0.2 prior standard deviations and θ being equal to 5 prior standard deviations. Figure 1 graphs the Probability that Modeler 1 is "Honest" Cond. on M_1 begin the "true" model as a function of r, the fraction of observations that the policy maker makes available to the modelers. For $\theta = 5/\sqrt{n\lambda^2}$ (right panel) the sample provides strong evidence that the restriction $\theta = \beta_1$ is satisfied. Thus, the posterior probability of the original model M_1 exceeds the posterior probability of the more flexible reference model M_{10} for almost all samples, regardless of r.

If $\theta = 0.2/\sqrt{n\lambda^2}$, most of the variation in the outcome y_i is due to the regression error u_i instead of the regressor $\theta(x_{i,j} + s_i)$. Thus, it becomes more difficult do distinguish the restricted model M_1 from the unrestricted model M_{10}. Due to the choice of prior the Modeler 1 chooses M_1 in the absence of any information, that is, $r = 0$. If $\tau = \tau_{min}$ then for $r \leq 0.5$ the sample contains no data from the treatment group, which makes it impossible to gather evidence against the restriction $\theta = \beta_1$ and the probability that M_1 is selected remains equal to one. Only for $r > 0.5$ does the modeler have data from both the control and treatment group and, in that case, the modeler reports results from M_{10} for about 25% of the samples. If $\tau = 0.5$ then the probability that Modeler 1 reports results from M_1 drops more quickly as a function of r and reaches the 75% level for $r > 0.3$.

Figure 2 depicts the probability that Modeler 2 finds confirmation for the restriction $\theta = \beta_2$ if data are generated from M_1. As in the case of Modeler 1, if $r \leq 0.5$ and $\tau = \tau_{min}$ the sample contains no information about the restriction $\theta = \beta_2$ and Modeler 2 always reports results from the restricted model M_2. If $r > 0.5$, for both $\tau = \tau_{min}$ and $\tau = 0.5$ there is some probability that Modeler 2 finds his restricted specification rejected against the more general reference model M_{20}. This probability generally increases with r. For $\theta = 0.2/\sqrt{n\lambda^2}$ the probability that Modeler 2 is "honest" drops to 80%, whereas it drops to about 40% if $\theta = 5/\sqrt{n\lambda^2}$ and a much larger fraction of the variation in the outcome variable is due to the explanatory variables. A comparison of Figures 1 and 2 indicates that for the small value of θ both modelers find their restrictions rejected with approximately equal probability.

We proceed by averaging over θ with respect to the prior distribution instead of conditioning on particular values of θ. Figure 3 depicts integrated probabilities that Modelers 1 and 2 are honest. The probabilities in the left panel are computed conditional on M_1 being "true", whereas the probabilities in the right panel are obtained by simulating data from M_2. Given the setup of our simulation experiment the two panels are symmetric and we focus on the left panel, which essentially averages over the two panels depicted in Figures 1 and 2. If $r \leq 0.5$, then $\tau_{min} = 0$. Thus, both modelers have no information that allows them to test the restriction of their models. In turn, they are honest with probability 1. If $\tau = 0.5$, then even for small values of r the modelers find their restrictions rejected with some probability.

For large values of r the difference between $\tau = 0.5$ and $\tau = \tau_{min}$ vanishes as $\tau_{\min} \longrightarrow 0.5$. Finally, conditional on M_1, the probability that M_2 finds his model rejected is higher than that of M_1 and vice versa.

5.3 Policy Maker's Risk

Having examined the behavior of the modelers under the *holdout* mechanism, we now consider the risk of using holdout samples for the policy maker. Figure 4 plots the probability that the modelers are honest *and* that the predictive-density-based selection yields the highest posterior probability model as a function of r for $\tau = \tau_{min}$ and $\tau = 0.5$. First, consider the case $\theta = 5/\sqrt{n\lambda^2}$. Overall, the choice of $\tau = \tau_{min}$ dominates $\tau = 0.5$. The probability function has an inverted U-shape. For $r \approx 0.5$, the policy maker finds the highest probability model almost with certainty. We conjecture that for small r the analysis suffers from imprecise estimates of θ and a diffuse distribution for $\hat{\theta}_p$. Large values of r, on the other hand, yield short holdout samples which makes it more difficult to measure the predictive performance of M_1 versus M_2. Second, if $\theta = 0.2/\sqrt{n\lambda^2}$ then the policy maker finds the highest posterior probability model with at most probability 1/2. For $\tau = 0.5$ and $r < 0.5$ there is a visible effect of predictive data mining, i.e. the modelers report results from their reference models M_{j0} instead of M_j.

We now turn to the risk differential $\Delta(\hat{y}_*, \hat{y}_p)$ in (16). The models M_j imply that only one of the two characteristics of the individuals is relevant for the outcome y_i. For our illustration, we assume that the policy maker's objective is to predict the effect of a subsidy $s_i = s_*$ for an individual whose relevant characteristic is $x_{i,j} = \sqrt{2}$ and whose irrelevant characteristic is $x_{i,(-j)} = 0.2\sqrt{2}$. Recall that $\sqrt{2}$ is the variance of the characteristics within the population and 0.2 their correlation.[6] Figure 5 depicts the average values for $(\hat{y}_p - \hat{y}_*)^2$ conditional on data generated from M_1 with either $\theta = 0.2/\sqrt{n\lambda^2}$ or $\theta = 5/\sqrt{n\lambda^2}$. The results mirror the probability of the policy maker finding the highest posterior probability

[6] Our setup creates a penalty for making good predictions with the wrong model – because the prediction will be based on the incorrect regressor value.

model. For large values of θ the policy maker can with $r = 0.5$ and $\tau_{min} = 0$ obtain a risk differential that is essentially zero.

We also plot the risk attained under the *no-holdout* mechanism, which leads the modelers to engage in in-sample data mining (see Option 3 in Section 4.3). The risk attained under *no-holdout* is large for both small and large values of θ. We deduce that in our numerical illustration the policy maker is able to achieve a lower risk by using a holdout sample. Finally, we plot the integrated risk differential $\Delta(\hat{y}_*, \hat{y}_p)$ of the policy maker in Figure 6, averaging over θ with respect to its prior distribution. The preferred strategy from the policy maker's perspective is to set $\tau = 0$ and $r = 0.5$.

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.[7] 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

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

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.

In future work we are planning to generalize our numerical results, extend our work to the analysis of non-random holdout samples, and allow for the possibilities that none of the structural models are correctly specified. Finally, our current analysis does not capture an important positive aspect of specification searches: they tend to eliminate model specifications that are clearly not empirically viable.

References

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

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

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

Kaboski, J., and R. Townsend (2011): "A Structural Evaluation of a Large-Scale Quasi-Experimental Microfinance Initiative," *Econometrica*, 79(5), 1357–1406.

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.

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

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

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.

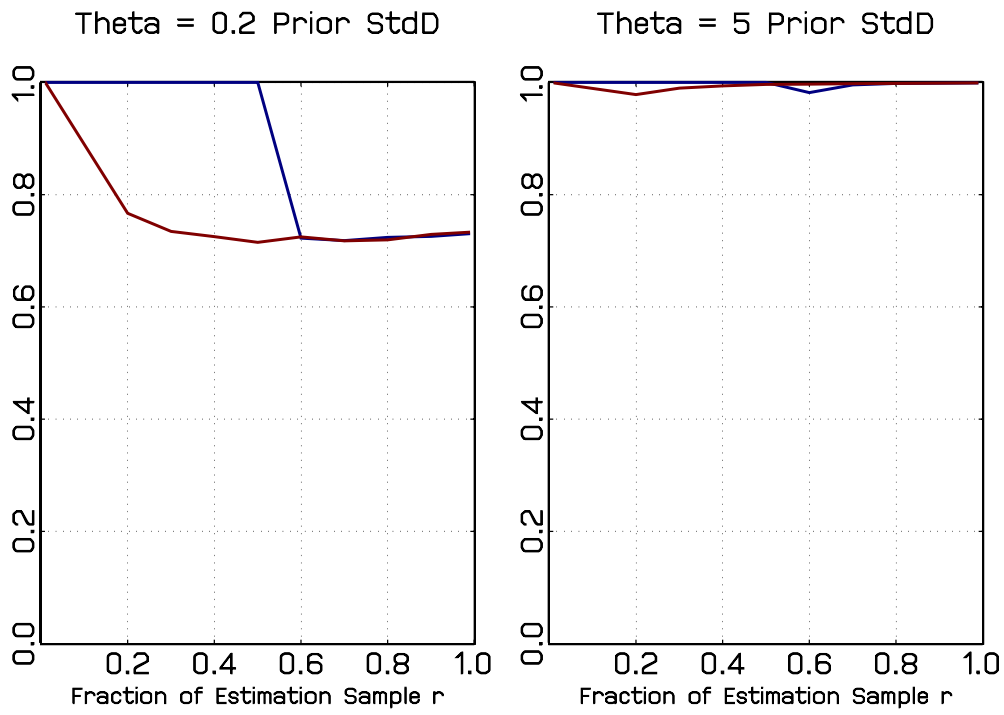
Winkler, R. (1969): "Scoring Rules and the Evaluation of Probability Assessors," *Journal of the American Statistical Association*, 64(327), 1073–1078.

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.

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

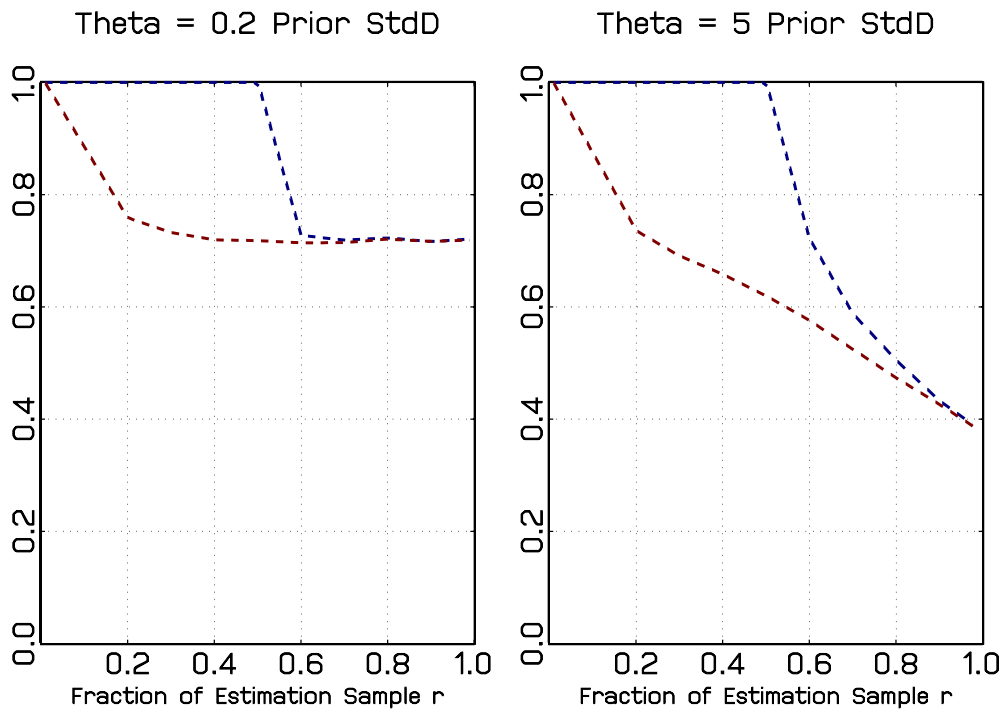
	$\tau = \tau_{min}$		$\tau = 0.5$	
	Control	Treatment	Control	Treatment
$r = 0.2$	200	0	100	100
$r = 0.5$	500	0	250	250
$r = 0.8$	500	300	400	400
$r = 1.0$	500	500	500	500

Figure 1: Probability that Modeler 1 is "Honest" Cond. on M_1



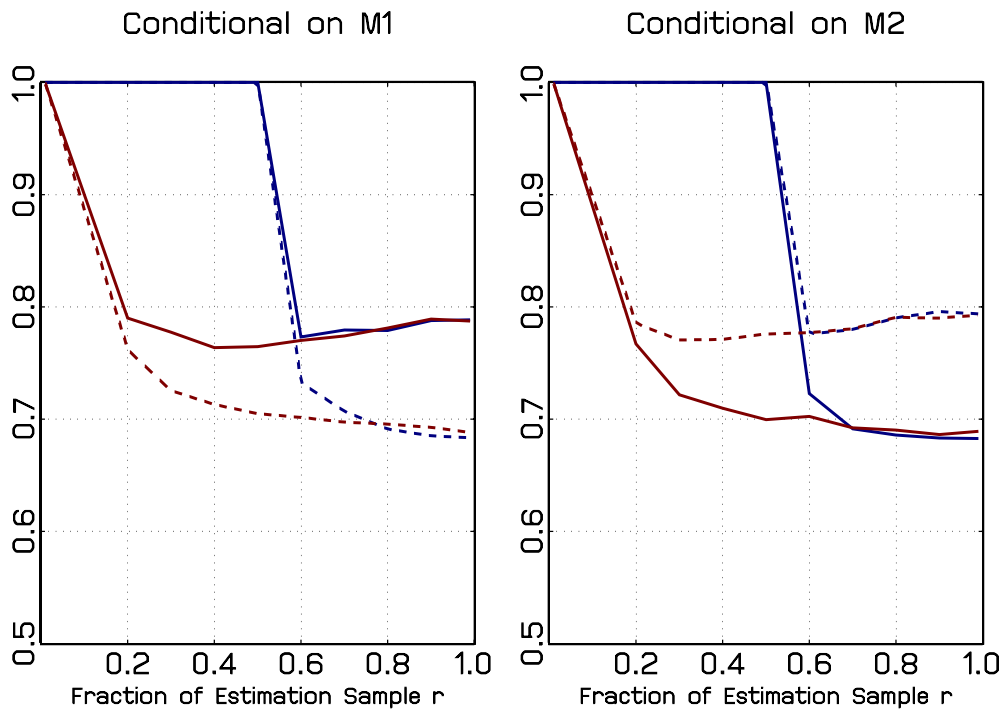
Notes: $\tau = \tau_{min}$ is blue, $\tau = 0.5$ is red.

Figure 2: Probability that Modeler 2 is "Honest" Cond. on M_1



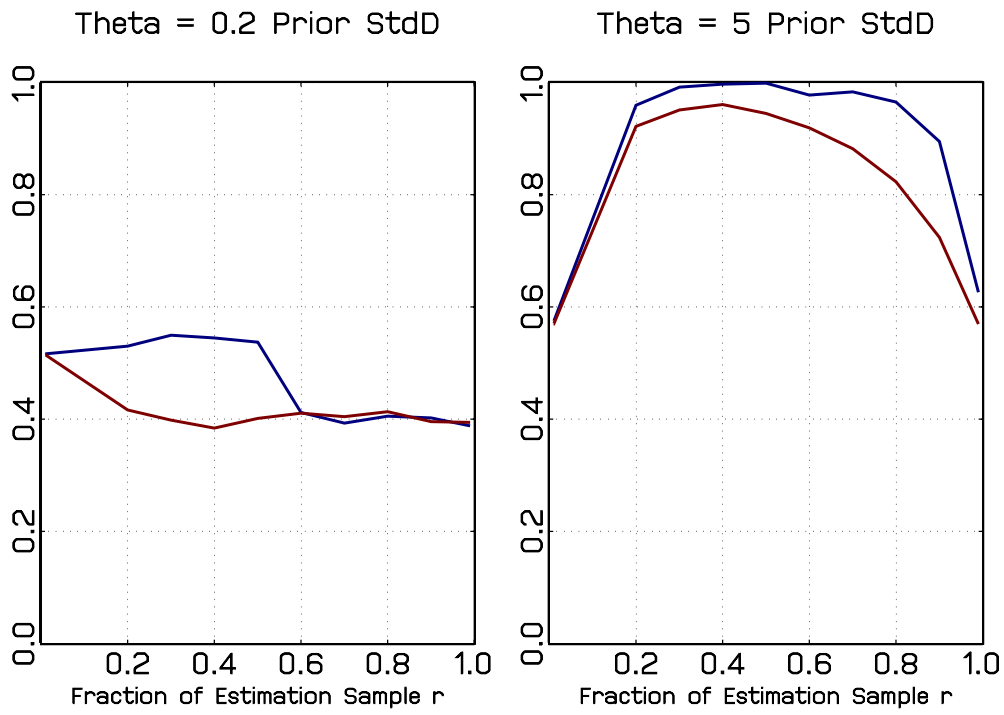
Notes: $\tau = \tau_{min}$ is blue, $\tau = 0.5$ is red.

Figure 3: Integrated Probability that Modelers are "Honest"



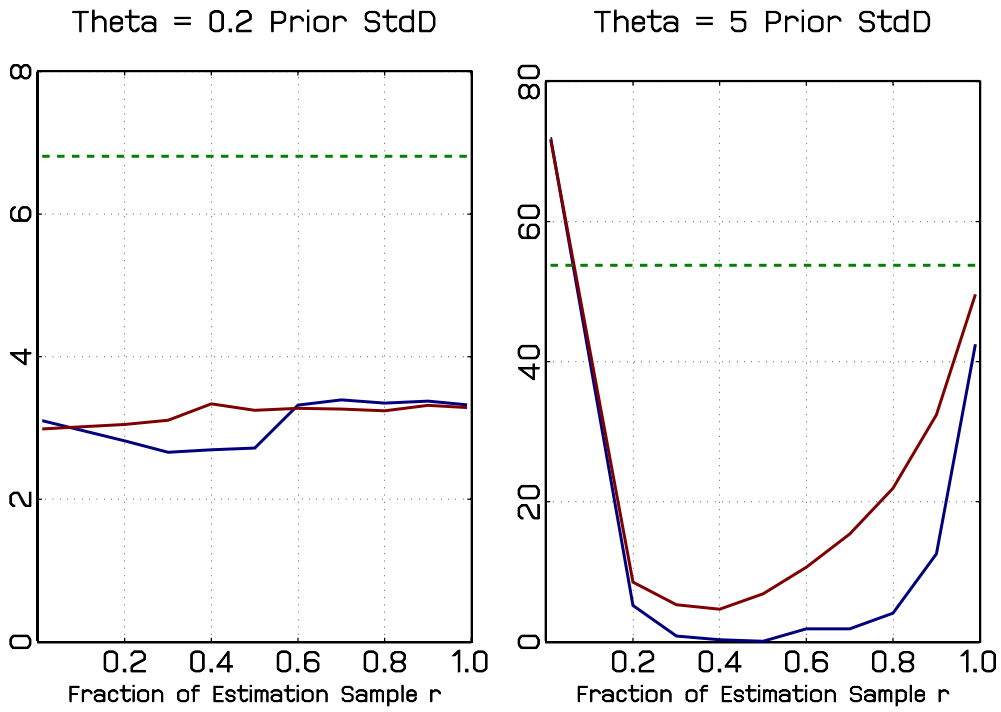
Notes: M_1 is solid, M_2 is dashed, $\tau = \tau_{min}$ is blue, $\tau = 0.5$ is red.

Figure 4: Prob. PM Finds Best Model Cond. on M_1 and θ



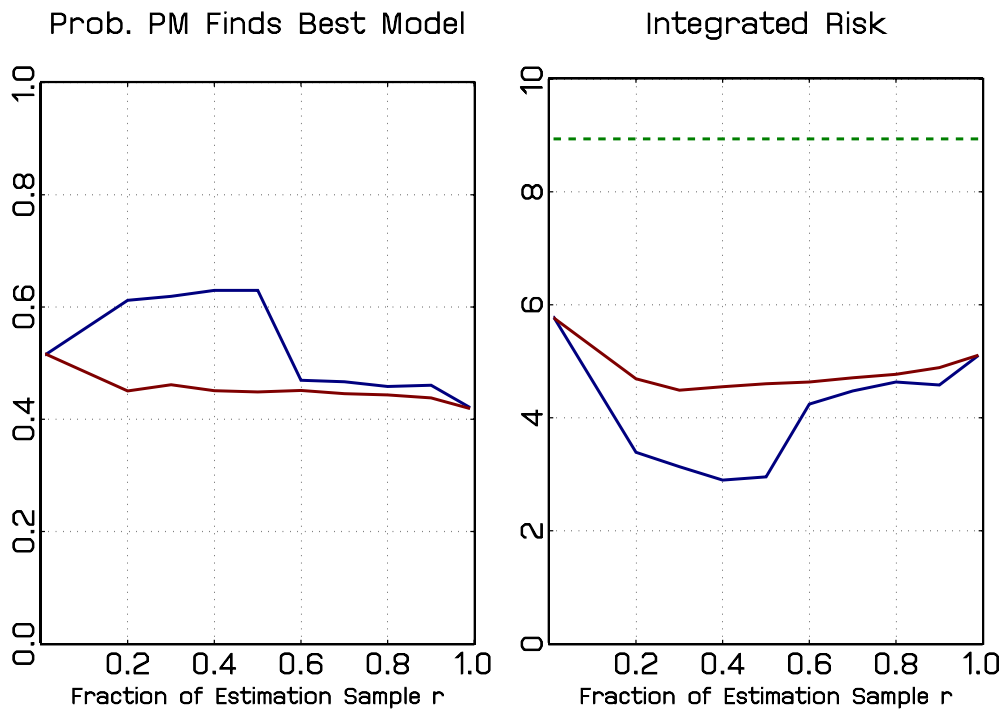
Notes: $\hat{\theta}$-density-based selection, $\tau = \tau_{min}$ is blue, $\tau = 0.5$ is red.

Figure 5: Risk Cond. on M_1 and θ



Notes: $\hat{\theta}$-density-based selection, $\tau = \tau_{min}$ is blue, $\tau = 0.5$ is red, Data mining on full sample is green.

Figure 6: Integrated Probability that PM Finds Best Model and Risk



Notes: $\hat{\theta}$-density-based selection, $\tau = \tau_{min}$ is blue, $\tau = 0.5$ is red, Data mining on full sample is green.

A Bayesian Analysis of the Linear Regression Model

Consider the linear Gaussian regression model, written in matrix form:

$$Y = X\theta + U. \tag{A.1}$$

The dimension of θ is $k \times 1$. Our examples focus on inference about θ, and we assume that the elements of U are $iidN(0,1)$. The likelihood function takes the form

$$p(Y|X,\theta) = (2\pi)^{-n/2} \exp\left\{(Y - X\theta)'(Y - X\theta)\right\}. \tag{A.2}$$

The prior for θ is normal, centered at 0 with covariance matrix $\underline{V}$. Thus, the prior density takes the form

$$p(\theta) = (2\pi)^{-k/2}|\underline{V}|^{-1/2} \exp\left\{-\frac{1}{2}\theta'\underline{V}^{-1}\theta\right\}. \tag{A.3}$$

According to Bayes Theorem the posterior distribution of θ is proportional to the product of prior density and likelihood function

$$p(\theta|Y,X) \propto p(\theta)p(Y|X,\theta).$$

The right-hand-side is given by

$$\begin{aligned} & p(\theta)p(Y|X,\theta) \\ & \quad \propto \quad (2\pi)^{-\frac{n+k}{2}}|\underline{V}|^{-1/2} \exp\left\{-\frac{1}{2}[Y'Y - \theta'X'Y - Y'X\theta - \theta'X'X\theta - \theta'\underline{V}^{-1}\theta]\right\}. \end{aligned}$$

The exponential term can be rewritten as follows

$$\begin{aligned} & Y'Y - \theta'X'Y - Y'X\theta - \theta'X'X\theta - \theta'\underline{V}^{-1}\theta \\ & \quad = \quad Y'Y - \theta'X'Y - Y'X\theta + \theta'(X'X + \underline{V}^{-1})\theta \\ & \quad = \quad \left(\theta - (X'X + \underline{V}^{-1})^{-1}X'Y\right)'\left(X'X + \underline{V}^{-1}\right)\left(\theta - (X'X + \underline{V}^{-1})^{-1}X'Y\right) \\ & \qquad +Y'Y - Y'X(X'X + \underline{V}^{-1})^{-1}X'Y. \end{aligned}$$

Thus, the exponential term is a quadratic function of θ. This information suffices to deduce that the posterior distribution of θ must be a multivariate normal distribution

$$\theta|Y,X \sim \mathcal{N}(\bar{\theta},\bar{V}) \tag{A.4}$$

with mean and covariance

$$\begin{aligned}\bar{\theta} &= (X'X + \underline{V}^{-1})^{-1}X'Y \\ \bar{V} &= (X'X + \underline{V}^{-1})^{-1}.\end{aligned}$$

In order to obtain the marginal likelihood, note that Bayes Theorem can be rewritten as follows

$$p(Y|X) = \frac{p(Y|X,\theta)p(\theta)}{p(\theta|Y,X)}.$$

Since, we previously showed that the posterior $p(\theta|Y,X)$ is multivariate normal all the terms on the right-hand-side are known:

$$\begin{aligned}p(Y|X) &= \frac{(2\pi)^{-n/2}(2\pi)^{-k/2}|\underline{V}|^{-1/2}\exp\left\{-\frac{1}{2}[(\theta-\bar{\theta})'\bar{V}^{-1}(\theta-\bar{\theta})]\right\}}{(2\pi)^{-k/2}|X'X+\underline{V}^{-1}|^{1/2}\exp\left\{-\frac{1}{2}[(\theta-\bar{\theta})'\bar{V}^{-1}(\theta-\bar{\theta})]\right\}} \\ &\quad \times \exp\left\{-\frac{1}{2}[Y'Y - Y'X(X'X+\underline{V}^{-1})^{-1}X'Y]\right\} \\ &= (2\pi)^{-n/2}|\underline{V}|^{-1/2}|X'X+\underline{V}^{-1}|^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[Y'Y - Y'X(X'X+\underline{V}^{-1})^{-1}X'Y]\right\}.\end{aligned} \tag{A.5}$$