# On the Use of Holdout Samples for Model Selection

*By* Frank Schorfheide and Kenneth I. Wolpin*

Researchers often hold out data from the estimation of econometric models to use for external validation. In this paper we examine possible rationales for this practice. For concreteness, two examples are considered. The first example (Todd and Wolpin (2008)) is taken from the microeconometrics literature. Suppose the goal is to evaluate the impact of a monetary subsidy to low-income households based on school attendance. A social experiment is conducted in which a randomly selected treatment sample of households is offered a school attendance subsidy at some level $\bar{s}$, whereas no subsidy is provided to the households in the control sample. In order to determine the optimal subsidy level, it is necessary to extrapolate the subsidy effect to other treatment levels. This requires the development and estimation of models that embed behavioral and statistical assumptions (structural models).

A holdout approach to the selection among competing structural models amounts to splitting a sample $Y$ into two subsamples, $Y_e$ and $Y_{ho}$. The models are then estimated based on $Y_e$, say the data from the control group, and ranked based on their ability to predict features of the holdout sample $Y_{ho}$, for instance the subsidy effect on the treatment group. Examples of this kind of external model validation in the context of randomized controlled trials are Wise (1985), Todd and Wolpin (2006), and Duflo, Hanna and Ryan (2011).

The second example is taken from the macroeconometrics literature. In time series analysis, competing models are often ranked in terms of their performance in pseudo-out-of-sample forecast comparisons. We will use the notation $Y_{n_1:n_2} = \{y_{n_1}, y_{n_1+1}, \ldots, y_{n_2}\}$, $n_1 \leq n_2$, with the understanding that $Y = Y_{1:n}$. Pseudo-out-of-sample forecasting performance is assessed by estimating the model based on $Y_{1:r}$, $r < n$, computing a forecast for $y_{r+1}$ and assessing the forecast error loss associated with this forecast. $y_{r+1}$ can be viewed as a holdout observation. In a recursive evaluation scheme this calculation is repeated for $r = n_0, n_0 + 1, \ldots, n - 1$ and the average forecast error loss is used as a measure of model performance.

Because it is not possible to conduct controlled experiments to determine the effects of macroeconomic policy changes, e.g. unexpected changes in the nominal interest rate, pseudo-out-of-sample forecast performance is also often used to rank macroeconomic models for the purpose of policy analysis. Policy predictions of models that forecast well (in our pseudo-out-of-sample sense) tend to be deemed more reliable than policy predictions of models that forecast poorly.[1]

For the remainder of the paper, we cast the problem of model selection in a Bayesian framework. Since the Bayesian approach allows us to assign probabilities to models, it provides a convenient analytical framework to study model uncertainty. In Section I we show that from a Bayesian perspective the use of holdout samples is suboptimal because the computation of posterior probabilities should be based on the entire sample and not just on a subsample. In Section II we argue that data-inspired mod-

* Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6297, USA. Schorfheide: schorf@ssc.upenn.edu. Wolpin: wolpink@ssc.upenn.edu. We thank Frank Diebold, George Mailath, Chris Sims, and participants of the session on "Model Selection" of the 2012 AEA Meetings for helpful comments and suggestions.

[1] Although it is possible to come up with counterexamples in which the model that delivers best forecasts generates the least accurate policy prediction, e.g. Kocherlakota (2007), it is difficult to convince a policy maker that a model that is not able to predict macroeconomic outcomes under the prevailing policy generates accurate predictions of macroeconomic outcomes under alternative policies.

ifications of structural models, i.e. models that embed behavioral and statistical assumptions, are likely to lead to an exaggeration of model fit. The use of holdout samples can in principle set an incentive not to exaggerate model fit. However, it comes at a cost: model weights are based on the fit of competing models to a subsample instead of the full sample which implies a loss of information. Section III reports on some ongoing research that aims to quantify this trade-off in a stylized principle-agent framework. Finally, a conclusion and an out-of-sample forecast of future work are provided in Section IV.

## I. A Prototypical Bayesian Analysis

Consider two models $M_1$ and $M_2$. Let $\theta_j$ denote the parameter vector associated with model $j$, $p(\theta_j|M_j)$ be the prior distribution, and $p(Y|\theta_j, M_j)$ be the density of the data given $\theta_j$ (likelihood function). According to Bayes Theorem the posterior distribution of $\theta_j$ is given by

$$(1) \quad p(\theta_j|Y, M_j) = \frac{p(Y|\theta_j, M_j)p(\theta_j|M_j)}{p(Y|M_j)}.$$

The term in the numerator is called the marginal likelihood (or data density) and is given by

$$(2) \quad p(Y|M_j) = \int p(Y|\theta_j, M_j)p(\theta_j|M_j)d\theta_j.$$

Marginal likelihoods are used to update Bayesian model probabilities. Let $\pi_{j,0}$ be the prior probability of model $M_j$ and $\pi_{j,n}$ its posterior probability after observing a sample of size $n$. Then, the ratio of posterior model probabilities (odds) is given by

$$(3) \quad \frac{\pi_{1,n}}{\pi_{2,n}} = \frac{\pi_{1,0}}{\pi_{2,0}} \frac{p(Y|M_1)}{p(Y|M_2)}.$$

The marginal likelihood can be rewritten in two ways, which are helpful for the interpretation of the statistic. First, consider the widely-used Schwarz (1978) approximation

$$\begin{aligned} p(Y|M_j) &\approx p(Y|\hat{\theta}_j, M_j) \\ (4) \qquad & \quad \times \exp\left\{ -\frac{\dim(\theta_j)}{2} \ln n \right\}, \end{aligned}$$

where $\hat{\theta}_j$ is the maximum likelihood estimate, $\dim(\theta_j)$ is the dimension of the model (typically the number of parameters), and $n$ is the sample size. (4) highlights that the marginal likelihood penalizes in-sample fit, that is, the value of the maximized likelihood function, by a measure of model complexity. Second, using the notation $Y_{n_1:n_2} = \{y_{n_1}, y_{n_1+1}, \ldots, y_{n_2}\}$ we can factorize the joint density of $Y = Y_{1:n}$ into a product of conditional densities

$$(5) \quad p(Y|M_j) = \prod_{i=1}^{n} p(y_i|Y_{1:i-1}, M_j),$$

where

$$\begin{aligned} p(y_i|Y_{1:i-1}) &= \int p(y_i|\theta_j, Y_{1:i-1}, M_j) \\ (6) \qquad & \quad \times p(\theta_j|Y_{1:i-1}, M_j)d\theta_j. \end{aligned}$$

Here the conditional density $p(y_i|Y_{1:i-1}, M_j)$ in (5), evaluated at the observed value, can be interpreted as a score of how well model $M_j$ is able to predict observation $y_i$ conditional on the information set $Y_{1:i-1}$. The subsequent representation in (6) highlights that $p(y_i|Y_{1:i-1})$ is obtained by first estimating $\theta_j$ based on $Y_{1:i-1}$ and then predicting $y_i$ conditional on this estimate.

Taking (3), (4), and (5), we reach the following conclusions. First, Bayesian model weights embody a specific measure of recursive out-of-sample fit. Unlike the hold-out schemes discussed in the introduction, the fit pertains to the entire sample $Y_{1:n}$ instead of just to the holdout sample. Second, the weights for each model are computed in a way that trades-off in-sample model fit with model complexity, penalizing overfitting. A frequently-cited reason for the use of hold-out samples is to discourage over-fitting of the data. The posterior model probabilities carry an automatic penalty for over-fitting.

## II. Data Mining and Specification Searches

The Bayesian analysis described in the previous section abstracts from two aspects of reality. First, empirical analysis in economics typically involves an incomplete model space. Thus, the investigator is aware that in addition to $M_1$ and $M_2$ there exist other plausible models for $Y$, but due to a resource constraint he is only able to analyze the former. Second, the specification of structural models is often the outcome of a lengthy development process that involves the careful inspection of a data set. Various approaches of weighting models and conducting policy analysis if the model space is incomplete are discussed, for instance, in Schorfheide (2000), Geweke (2010), and Geweke and Amisano (2012). In the remainder of this paper, we focus on the second issue.

Developing and estimating structural micro- or macroeconomic models is a lengthy, tedious, and time-consuming endeavor. We refer to the process by which a modeler tries to improve the fit of a structural model during the model development process as *data mining.* This process might involve changes in functional forms, the introduction of latent variables, or the inclusion of particular covariates. A good example of a data-mined model in the literature on dynamic stochastic general equilibrium (DSGE) macroeconomic models – and we do not mean this as a criticism – is the widely-used Smets and Wouters (2007) model.

The core of the Smets-Wouters model is comprised of a neoclassical stochastic growth model. It is augmented by mechanisms that generate price and wage stickiness, which in turn generate a role for monetary policy. Many researchers contributed to the development of the model. Most notably, Christiano, Eichenbaum and Evans (2005) augmented a prototypical New Keynesian DSGE model that was emerging in the late 1990s by habit formation and adjustment cost mechanisms that allowed the model to match the impulse responses of a large set of variables to a monetary policy shock, as measured with a structural vector autoregression.

Building on the Christiano, Eichenbaum, and Evans specification, Smets and Wouters (2003) introduced additional exogenous disturbances so that the resulting DSGE model could track seven key macroeconomic variables, including GDP, inflation, and interest rates, for the Euro Area. In subsequent work, the model was refined further by changing the law of motion of the exogenous shocks and making some small modifications to the endogenous propagation mechanism. Based on these refinements, the model has been able to forecast U.S. macroeconomic variables, with the exception of the 2008-09 recession, quite well in comparison with other macroeconometric models. In its final form the Smets and Wouters (2007) model is the outcome of a sequence of data-based modifications of some initial New Keynesian DSGE model, that is, of mining.

On the positive side, the data mining process has led from a fairly restrictive neoclassical stochastic growth model that had severe problems capturing salient features of the data, to the discovery of a much more elaborate specification that appears to be competitive in its ability to track and forecast macroeconomic time series with less restrictive vector autoregressive models.

On the negative side, the outcome of data mining might be that results from versions of the model that only fit slightly worse than the best specification are forgotten and do not get reported in the final analysis. Thus, as part of the data mining, the prior distribution is essentially shifted toward a model specification that fits the data well. Abstracting from the decomposition of a DSGE model into sub-models, this process is akin to changing a prior from $p(\theta)$ to $\tilde{p}(\theta)$, whereby $\tilde{p}(\theta)$ assigns high prior probability to the region of the parameter space that is associated with a high likelihood function $p(Y|\theta)$. In turn, the marginal likelihood computed based on $\tilde{p}(\theta)$ overstates the fit of the reported DSGE model and the posterior distribution understates the parameter uncertainty. This negative aspect of data mining can provide a rationale for the explicit use of holdout samples in model selection.

## III. Can Holdout Samples Provide a Remedy?

In ongoing research we are developing a principle-agent framework in which data-mining generates an impediment for the implementation of the ideal Bayesian analysis (Schorfheide and Wolpin (2011)). The framework in its current state is tailored toward the treatment effect example discussed in the introduction. The gist of our analysis is the following.

A policy maker conducts a social experiment in order to measure the effectiveness of a subsidy. A randomly selected sample receives the subsidy and a second randomly-selected sample of equal size does not receive the subsidy. In order to extrapolate the treatment effect to other levels of the subsidy, the policy maker engages two modelers who each can fit a structural model to whatever data they receive from the policy maker and provide predictions of the treatment effect. The structural models embody enough restrictions such that policy predictions can be generated even in the absence of any information from the treatment sample.

The modelers are rewarded based on the fit of the model that they are reporting to the policy maker. This creates an incentive to engage in the kind of data mining described in Section II. As explained in the context of the Smets-Wouters model, the data mining process is likely to raise the marginal likelihood because the contribution of less-likely versions of the structural model to the marginal likelihood are eliminated during the specification search.

What are the potential benefits of a holdout sample? Suppose that the policy maker splits the sample $Y$ into an estimation sample $Y_e$ and a holdout sample $Y_{ho}$. The modelers have access to $Y_e$ in order to estimate their models and have to provide a predictive density for $Y_{ho}$. If the modelers are rewarded according to a log-scoring rule that makes their payoff proportional to the log of the reported predictive density for $Y_e$, then they have an incentive to truthfully reveal their subjective beliefs, which in our case correspond to $p(Y_{ho}|Y_e, M_j)$. This insight dates at least back to Winkler (1969).

On the positive side, the holdout mechanism discourages the modelers from overly down-weighting regions of the parameter space that make the observed data appear unlikely. On the negative side, the policy maker will base the weights assigned to models $M_1$ and $M_2$ at best on the ratio

$$\frac{p(Y_{ho}|Y_e, M_1)}{p(Y_{ho}|Y_e, M_2)} = \frac{p(Y_{ho}, Y_e|M_1)}{p(Y_{ho}, Y_e|M_2)} \Big/ \frac{p(Y_e|M_1)}{p(Y_e|M_2)}, \tag{7}$$

ignoring the relative fit of the two models on the estimation sample $Y_e$. Because the first-best Bayesian analysis described in Section I is not implementable by the policy maker, the question becomes as to what composition of the hold-out sample allows the policy maker to generate model weights that are close to the ideal Bayesian model weights. In Schorfheide and Wolpin (2011) we report results from simulation experiments that suggest providing the modelers only with data from the treatment group.

## IV. Conclusion and Outlook

Two important consequences of data mining (in the sense that this term is used in this paper) are a distortion of weights of competing models and an under-statement of uncertainty. Both have adverse effects on the prediction of future events and the design of economic policies. In principle, holdout samples can be used to set incentives for modelers to truthfully reveal their findings and not to understate uncertainty. In practice there is, however, a question of implementation.

In the context of the treatment effect example, it is conceivable that the policy maker who conducts the social experiment actually withholds observations from academic consultants who are tasked to develop structural models for the extrapolation of treatment effects.

In the context of comparing macroeconomic models the goal is typically to make a forecast or policy decision at a particular point in time, say $T$. In period $T$ each modeler technically has full access to all the historical data. Reaping the incentive benefits from the use of holdout samples would require a commitment from the modelers not to look at part of the data even though they are stored on their computers.

## REFERENCES

**Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans.** 2005. "Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy." *Journal of Political Economy*, 113(1): 1–45.

**Duflo, Ester, Rema Hanna, and Stephen Ryan.** 2011. "Incentives Work: Getting Teachers to Come to School." *American Economic Review*, forthcoming.

**Geweke, John.** 2010. *Complete and Incomplete Econometric Models.* Princeton University Press, Princeton.

**Geweke, John, and Gianni Amisano.** 2012. "Prediction with Misspecified Models." *American Economic Review (Papers and Proceedings)*, 102(2).

**Kocherlakota, Narayana R.** 2007. "Model Fit and Model Selection." *Federal Reserve Bank of St. Louis Review*, 89(4): 349–360.

**Schorfheide, Frank.** 2000. "Loss Function-based Evaluation of DSGE Model." *Journal of Applied Econometrics*, 15(6): 645–670.

**Schorfheide, Frank, and Kenneth Wolpin.** 2011. "To Hold Out or Not to Hold Out." *Manuscript, http://www.ssc.upenn.edu/˜schorf/.*

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

**Smets, Frank, and Rafael Wouters.** 2003. "An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area." *Journal of the European Economic Association*, 1(5): 1123–1175.

**Smets, Frank, and Rafael Wouters.** 2007. "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review*, 97(3): 586–606.

**Todd, Petra, and Kenneth 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.

**Todd, Petra, and Kenneth Wolpin.** 2008. "Ex Ante Evaluation of Social Programs." *Annales d'Economie et de Statistique*, 91-92: 263–292.

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

**Wise, David.** 1985. "Behavioral Model versus Experimentation: The Effects of Housing Subsidies on Rent." In *Methods of Operations Research 50.* , ed. Peter Brucker and R. Pauly, 441–489. Königstein: Verlag Anton Hain.