Avoiding biases from data-dependent specification search: an application to a tillage choice model

Authors
Sanchita Sengupta, Lyubov A. Kurkalova, and Catherine L. Kling[1]

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006

[1] Sengupta is graduate assistant, Center for Agricultural and Rural Development at Iowa State University, Kurkalova is Assistant Professor at the Department of Agribusiness Economics, Southern Illinois University, and Kling is Professor of Economics and Head, Resource and Environmental Policy Division of the Center for Agricultural and Rural Development at Iowa State University.

Avoiding biases from data-dependent specification search: an application to a tillage choice model*

Sanchita Sengupta, Lyubov A. Kurkalova, and Catherine L. Kling†

March 17, 2006

Abstract

The study evaluates the gains of avoiding data-dependent specification search on an estimation sample in an application to discrete choice models. We incorporate data splitting, the process by which the total available sample is randomly split in two or more sub-samples with the first (specification) sub-sample used for specification search, and the second (estimation) sub-sample used for obtaining "clean' estimates using the model chosen on the specification sub-sample according to a set criterion. We estimate 14 binary Logit models of the adoption of conservation tillage corresponding to the major sub-watersheds of the Upper Mississippi River Basin. For each of the sub-watershed models, we use the specification sub-sample to choose the explanatory variables that lead to the highest number of correct predictions provided that estimated coefficients are in conformity with economic theory. To evaluate the gains of avoiding specification search on the estimation sub-sample, we follow Gong (1986)[8] and calculate the expected excess error, which is a measure of excess optimism concerning

*Draft, please do not site without the authors' permission

†Sengupta is graduate assistant, Center for Agricultural and Rural Development at Iowa State University, Kurkalova is Assistant Professor at the Department of Agribusiness Economics, Southern Illinois University, and Kling is Professor of Economics and Head, Resource and Environmental Policy Division of the Center for Agricultural and Rural Development at Iowa State University.

model fit on the specification sample. We find that the excess optimism varies with the sub-watersheds and has a tendency to be larger for the sub-watersheds with smaller samples.

1 Introduction

Estimation of econometric model parameters customarily assumes that the model structure is known. However, economic theory oftentimes provides only a partial guidance on the model structure, leaving the choice of the model's functional form and/or the set of explanatory variables to the researchers. This model uncertainty then leads to specification search by which explanatory variables are selected into the model to provide the best model specification according to preset criteria. However, if the same sample is used for both selecting the model and for fitting the model and making inferences, too narrow prediction intervals and biases in parameter estimates can ensue (Chatfield, 1995). In consequence, coefficient estimates and standard errors following pretesting cannot be used for valid inference (Veall 1992[14], Potscher 1991 [13]). Although the presence of non-trivial biases that result from data-dependent specification search is widely recognized by statisticians (Chatfield (1995) [1], Leamer (1983) [10]), it is rarely taken into account in applied econometrics. Some exceptions to this practice are Creel and Loomis (1990)[2] and Herriges et.al.(2005)[6], who take into consideration the bias in inferences that arise due to specification search.

Admittedly, model uncertainty is difficult to quantify. The commonly proposed remedial approaches include the Bayesian Model Averaging Approach, collection of more data, and data splitting (see, e.g., Chatfield(1995)[1]). This study focuses on data splitting, the process by which the total available sample is randomly split in two or more sub-samples with the first (specification) sub-sample used for specification search, and the second (estimation) subsample used for obtaining "clean" estimates using the model chosen on the specification sub-sample according to a set criterion. The other sub-samples (if any) are then used to further evaluate model fit. Since data sets available to researchers are almost never of the size permitting such procedure, this approach is rarely used in the applied work and the studies reporting specification search biases are similarly scarce. Our analysis aims at filling this gap by evaluating the excess optimism concerning model fit attributable to data-specification search on the estimation sample in an application to

discrete choice models.

In this paper we perform systematic data analysis and investigate the effects of data-dependent specification search for a data set that originally contains some 37,000 data points. We incorporate data splitting to estimate several binary logit models of the adoption of conservation tillage corresponding to major sub-watersheds of the Upper Mississippi River Basin, and estimate the excess optimism concerning model fit that is attributable to the data-specification search, using the approach developed by Gong (1986).

The rest of the paper is organized as follows. In section 2, we discuss why model uncertainty could be a problem and the different ways that have been used to deal with this problem. Section 3 presents an empirical application to the estimation of discrete choice models of conservation tillage adoption, and section 4 concludes.

2 Model uncertainty

Pretesting or preliminary testing of the data to determine the type of model that is likely to be applicable, is a potential problem in statistics. Pretesting could entail a coefficient restriction, testing for heteroscedasticity or serial correlation or as in our case, searching for the model with the largest number of correct predictions. Zhang(1992)[17] provide asymptotic results for inference after selecting a linear regression model based on final error prediction criterion. He finds the asymptotic variance to be satisfactory but asymptotic confidence regions to be too small. The problem is aggravated for small samples. But large sample with excessive data mining is also likely to lead to invalid inference. The Optimism Principle defined by Picard and Cook (1984)[12], that model fitting necessarily gives optimistic results, is a manifestation of model uncertainty.

There are two schools of thought on the approach to dealing with model uncertainty, Bayesian and frequentist. Bayesian Model Averaging requires taking the weighted average of candidate models. The weights used are the Bayesian posterior probabilities and since they depend on the specification of prior probabilities, they are difficult to compute especially where there is no true model. Further, if the population form is uncertain, computing the Bayes factor could be another problem. We employ a frequentist approach in this study.

In the spirit of scientific inference which 'involves collecting many sets of

data and establishing a relationship which generalizes to different conditions' (Chatfield, 1995), the ideal frequentist approach to solving model uncertainty is to use an existing data set for model selection through testing and then collect new data to estimate the selected model. However, collecting more data is expensive in most economic studies. A viable alternative to collection of new data to perform out-of-sample inference is data splitting.

2.1 Data splitting and model selection

According to Faraway(1998)[7], if a large data set is available, the best way to perform out-of-sample analysis is by a three-way random data split. The first set (specification set) should be used for selection of model, the second (estimation set) for estimation of the parameters and for point prediction and the third (validation set) for assessing the variability of the predictions. However, Faraway (1998) has noted that 'the purpose of data splitting is to obtain better estimates of the variability of predictions, and the price one pays is that the actual variability of the predictions will tend to be higher' as the size of the estimation sample is smaller than that of the original sample.

An important step in model selection is the selection of a criteria. There is no universally acceptable model selection criteria in the discrete choice models, but two common approaches are to select models with largest value of pseudo R^2 and the largest number of correct predictions (Veall and Zimmermann, 1996 [15]). The goodness-of-fit statistic that is used in this study for specification search is the "percent correctly predicted". Specifically, we assume that a choice is correctly predicted if the predicted probability of the choice is greater or equal to 0.5. The threshold of 0.5 is not suitable for every discrete choice model (see, e.g., a discussion in Norwood et al., 2004[11]), but it works in our situation, since, as it will be clear from the application below, the cost of misclassifying one alternative is not very different from the cost of misclassifying the other alternative. In this paper, we first split the data set applying the algorithm suggested by Faraway (1998) and choose the best fitting model based mostly on the goodness-of-fit criterion. We then use bootstrap methods to assess the benefits of avoiding specification search on the estimation sample.

2.2 Bootstrap methods for estimating excess optimism

To estimate the excess optimism concerning model fit that is attributable to the data-dependent specification search, we employ bootstrap (resampling) techniques originally developed to correct for the optimism when data splitting is not an option (Efron (1982)[3], Efron and Gong (1983)[4], and Efron and Tibshirani (1993)[5]). As Efron and Gong (1983) point out, although theoretical basis for these methods is limited, the techniques can be successfully used in practice. The methods are based on the assumption that the original data set represents the underlying population and random draws from the original sample are draws from the same population.

The estimation of the excess optimism is based on the following observation (Efron,1982). Since the criteria for selecting the binary choice model with the best fit is the largest number of correct predictions, the prediction error or the apparent error is the number of incorrect predictions. Thus, the model selection bias can be manifested in the optimistic value of this apparent error. We follow Gong (1986) who proposed bootstrap methods to estimate the expected excess error.

3 Application

Agriculture in the Midwest has been targeted for conservation practices by various federal and state incentive-based programs. To better estimate the costs of current and intended programs and to better target conservation program expenditures there is an imperative need to understand the farm-level costs of conservation practices adoption for large, diverse areas. This study estimates these costs for one of the most effective conservation practices, conservation tillage (CT), for the entire Upper Mississippi River Basin (UMRB), an area which encompasses parts of Iowa, Illinois, Missouri, Wisconsin and Minnesota. The methodology we apply builds upon the work of Kurkalova et al. (2006)[9] who estimate the costs of CT adoption for the state of Iowa.

3.1 Study region and data

The study region, the Upper Mississippi River basin (UMRB) is defined as U.S. Geological Survey hydrologic region 07 (http://water.usgs.gov). UMRB covers 492,000 square kilometers in parts of Iowa, Illinois, Missouri, Wisconsin and Minnesota. The entire basin is divided into sub-watersheds or 4-digit

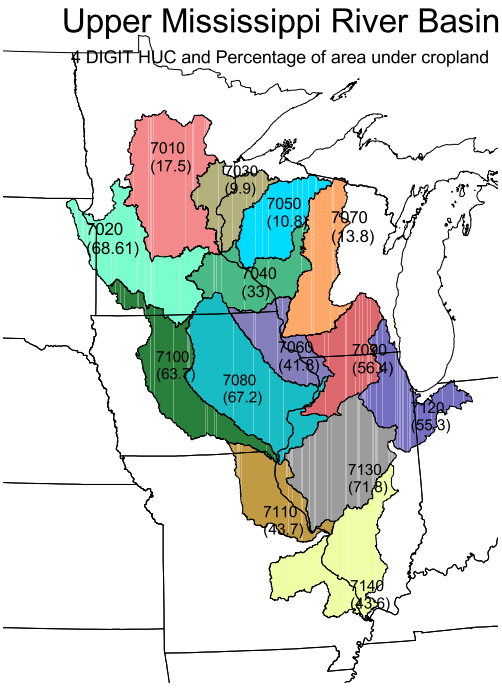

Figure 1: 4 digit Hydrologic Units in the Upper Mississippi River Basin

hydrologic units (HUC) that indicate the hydrologic region (first two digits) and hydrologic subregion (second two digits). There is substantial heterogeneity across the UMRB in terms of land use. As it can be seen from Figure 1, the percentage area that is under cropland ranges from a minimum of 9.9% in HUC 7030 to 68% to HUC 7020. Incidentally, the major parts of both of these HUCs are in Minnesota. To reflect this heterogeneity, we estimate several CT adoption models corresponding to the sub-watersheds.

The data comes primarily from the Natural Resource Inventory (NRI) (Nusser and Goebel, 1997, USDA/NRCS, 1994). The NRI is a scientifically based, longitudinal panel survey of soil, water, and related resources, designed to assess conditions and trends every five years. The 1997 NRI provides results that are nationally consistent for all nonfederal lands for four points in time 1982, 1987, 1992, and 1997. However, conservation tillage information is provided only in 1992 and hence only the 1992 data set is used for this study. The NRI data set for the UMRB region consists of a

total of 103,849 observations. Table 1 shows the distribution of these points across the 4-digit HUCs and under corn, soybean production and conservation tillage. Most of the UMRB area is under corn production. In consent with climatic conditions, the northern HUCs have fewer soybean acres than the southern HUCs and tillage adoption is higher in the south than in the north. The NRI data set further provides information on geo-physical properties of the land, i.e. soil characteristics, slope, erodibility, and the like. The complete data set is formed by adding constructed net returns, climatic data and farm characteristics as in Kurkalova et al. (2006).

The economic theory provides a guidance only on which groups of variables ought to be present in the set of explanatory variables (such as the crop grown, soil and landscape characteristics of cropland, farmer characteristics, and climatic variables), and for the sake of brevity, we refer interested readers to Kurkalova et al. (2006) for the details on the rationale for each of the groups of the variables. Table 2 provides variable descriptions and summary statistics for the combined data set.

3.2 Adoption models

The models that are similar to that of Kurkalova et al. (2006) are derived under the assumption that a farmer adopts conservation tillage if the expected annual net returns from this farming practice, π_1, exceed those from the alternative, conventional tillage, π_0, plus a premium, P, associated with uncertainty. Then, assuming that $\pi_1 - P$ is a linear function of a set of observed explanatory variables x and that the observations on π_0 are available, the model is given by

$$\Pr[Y = 1] = \Pr[\pi_1 \geq \pi_0 + P + \sigma\epsilon] = \Pr\left[\epsilon \leq \frac{\beta' x}{\sigma} - \frac{\pi_0}{\sigma}\right], \qquad (1)$$

where ϵ is a logistic error and the observed dependent variable Y takes on the value of 1 of CT is adopted and zero otherwise. The parameters of interest are the linear function parameters β together with σ, the error term multiplier.

The specific models for each of the sub-watersheds are the variants of the

basic specification, where

$$\begin{aligned}\beta' x = {} & \beta_0 + \beta_{0,c} I_c + \beta_{0,s} I_s \\ & + \beta_1 SLOPE + \beta_2 PM + \beta_3 AWC \\ & + \beta_4 EI + \beta_5 OM + \beta_6 PH \\ & + \beta_7 TMAX + \beta_8 TMIN + \beta_9 PRECIP \\ & + \beta_{10} TENANT + \beta_{11} OFFARM + \beta_{12} AGE \\ & + \beta_{13} MALE + \beta_{14} CODE \\ & + PRSTD(\beta_{15} + \beta_{16}\pi_0 + \beta_{17} TENANT \\ & + \beta_{18} OFFARM + \beta_{19} AGE + \beta_{20} MALE \\ & + \beta_{21} CODE)\end{aligned}$$

In addition to the specification described above, we also consider a specification that describes the probability of adopting conservation tillage as a function of the difference in the net returns between conventional and conservation tillage. In this case, instead of viewing the returns to conventional tillage as being known and that to conservation tillage being unknown, it is assumed that the average returns to both tillage methods are known. In this case, the model can be written as

$$\Pr[Y = 1] = \Pr[\pi_1 \geq \pi_0 + P + \sigma\epsilon] = \Pr\left[\epsilon \leq \frac{\beta' x}{\sigma} - \frac{\pi_{0-1}}{\sigma}\right], \qquad (2)$$

where π_{0-1}denotes the difference in net returns to conventional and conservation tillage. In this specification, $\beta' x$ represents the negative of the risk premium, rather than the difference between the expected net returns from conservation tillage and the risk premium. We refer to models (1) and (2) as net returns (NR) and difference (D) models, respectively.

3.3 Results: specification search

To conduct specification search, we split each HUC's sample randomly in 4 sub-samples, and use the first, specification, sub-sample for specification search. In this search, we choose the specification that leads to the highest number of correct predictions, provided that the estimate of $1/\sigma$, which is the negative of the estimated coefficient of π_0 in the NR model and is the

negative of the estimated coefficient of π_{0-1} in the D model, is positive as required by the theory. This way, we find the best model structure, and then obtain specification-search-bias-free estimates for the chosen models on the second (estimation) sub-sample. We chose the best-fitting models by varying the following model specifications:

1. *Area*: for each HUC, we choose the contiguous area containing the HUC,
2. *Variable*: choice among different soil and farmer characteristics variables,
3. *Model*: choice between the NR and D models.

Table 3 provides parameter estimates and their standard errors after specification search. (on the estimation sample). Table 4 provides the percentages of correct predictions for the following four combinations of parameter estimates and data sets:

1. Specification sample and parameter
2. Estimation sample and parameter
3. Specification parameter and estimation sample
4. Estimation parameter and validation sample

3.4 Computing excess optimism

To estimate the excess optimism concerning model fit that is attributable to the data-specification search, we follow Gong (1986). Specifically, we consider the observed sample, $\mathbf{Z}_1 = (y_1, \mathbf{X}_1), ..., \mathbf{Z}_N = (y_N, \mathbf{X}_N)$ as being independent and identically distributed from an unknown distribution F. Here matrix $\mathbf{X}$ is defined as $\mathbf{X} = \begin{pmatrix} x \\ -\pi_0 \end{pmatrix}$ for the NR model, and as $\mathbf{X} = \begin{pmatrix} x \\ -\pi_{0-1} \end{pmatrix}$ for the D model. Let matrix β be defined as $\beta = \begin{pmatrix} \beta/\sigma \\ 1/\sigma \end{pmatrix}$ The prediction rule $\eta = \eta(\beta, \mathbf{X})$ associated with the model is the rule that allows predicting the value y_0 of the CT adoption indicator for any new set of observed explanatory

variables $\mathbf{X}_0$. Let $e_0 = \beta' \mathbf{X}_0$. The prediction rule η is given by the following: $y_0 = 1$, if $exp\left(e_0\right) / \left(1 + exp\left(e_0\right)\right) > 0.5$, and $y_0 = 0$ otherwise.

Let $Q\left(y_0, \eta\left(\beta, \mathbf{X}_0\right)\right)$ be the criterion that scores the discrepancy between the observed value y_0 and its predicted value $\eta = \eta\left(\beta, \mathbf{X}_0\right)$, which takes on the value of one if the observed and the predicted values are different, and zero otherwise. Let $\hat{F}$ be the empirical distribution function that puts mass $1/N$ at each point $\mathbf{Z}_1, ..., \mathbf{Z}_N$. The *true error* of is defined to be the expected error that the set of estimates makes on a new observation $\mathbf{Z}_0 = \left(y_0, \mathbf{X}_0\right)$ from F, $q = q\left(\hat{F}, F\right) = E_{z_0 \tilde{} F} Q\left(y_0, \eta\left(\beta, \mathbf{X}_0\right)\right)$. The *apparent error* of η is defined as $\hat{q}_{app} = q\left(\hat{F}, \hat{F}\right) = E_{z_0 \tilde{} \hat{F}} Q\left(y_0, \eta\left(\beta, \mathbf{X}_0\right)\right) = \frac{1}{N} \sum_{i=1}^{N} Q\left(y_i, \eta\left(\beta, \mathbf{X}_i\right)\right)$. Finally, the difference $R\left(\hat{F}, F\right) = q\left(\hat{F}, F\right) - q\left(\hat{F}, \hat{F}\right)$ is the *excess error*, and the expression $r = E_{\hat{F} \tilde{} F} R\left(\hat{F}, F\right)$ is the *expected excess error* of the prediction rule $\eta = \eta\left(\beta, \mathbf{X}\right)$. Here the expectation is taken over $\hat{F}$, which is obtained from $\mathbf{Z}_1, ..., \mathbf{Z}_N$ generated by F. If no data-dependent specification search has been conducted, than the expected excess error is zero. However, if data-dependent specification search has been performed, than the expected excess error is positive and thus is a reasonable measure of the excess optimism concerning model fit.

The bootstrapping procedure to compute the measure of optimism evolves in the following steps:

1. Let N be the number of observations in the sample $\mathbf{Z} = \{\mathbf{Z}_1, ..., \mathbf{Z}_N\}$. Take N random draws with replacement from $\mathbf{Z}$. These constitute one bootstrap sample, $\mathbf{Z}^b$. Estimate the selected logit model on the sample and obtain the bootstrap estimate $\hat{\beta_b}$.

2. Compute predicted probability with bootstrap estimates $\hat{\beta}_b$ and bootstrap sample explanatory variables $\mathbf{X}^b$ as $Y^*_{bi} = \frac{exp(\hat{\beta}_b \mathbf{X}^b_i)}{(1+exp(\hat{\beta}_b \mathbf{X}^b_i))}$ for $i = 1.....N$.

3. Compute predicted probability with bootstrap estimates $\hat{\beta}_b$ and the original sample $\mathbf{X}$ as $Y^*_{obi} = \frac{exp(\hat{\beta}_b \mathbf{X}_i)}{(1+exp(\hat{\beta}_b \mathbf{X}_i))}$ for $i = 1.....N$.

4. Apply the prediction rule η with the 0.5 threshold and obtain the proportion of incorrect predictions for both predicted probabilities,

$q_{b0} = \frac{1}{N}\sum_{i=1}^{N} Q_(bo)$ and $q_b = \frac{1}{N}\sum_{i=1}^{N} Q_(b)$, where Q_{bo} is estimated using Y^*_{obi} and Q_b is estimated using Y^*_{bi}.

5. Repeat 1, 2, 3 and 4 a large number B times.

6. Obtain the estimate of the expected excess error, which is the average of the difference between two proportions taken over all bootstrap samples as $\omega = \frac{1}{B}\sum_{b=1}^{B}[q_{b0} - q_b]$.

Table 5 reports the estimates of the average error and the distribution of the measure of optimism ω over 1,000 bootstrap samples, for 3 different watersheds, HUC 7080, HUC 7100, and HUC 7110 with 1,641, 856, and 412 observations in the specification data set, respectively. Somewhat surprisingly, we get little difference in the model fit between the specification and estimation samples. An average error of 0.33 for HUC 7080 means that 33% of the time we get wrong predictions with the specification sample, while with estimation sample we get wrong prediction 32% of the time. If we correct for the optimism by adding the expected excess error estimates to the apparent error rates we get the bias corrected estimates as 34% for the specification sample and 33.5% for estimation sample.

Excess error results from computing the difference between the average number of incorrect predictions using the original sample and the bootstrap estimates, and the average number of incorrect predictions using the bootstrap samples and bootstrap estimates. The mean value of the optimism measure is positive, indicating that the apparent error tends to underestimate the prediction error. The magnitude of optimism is small, indicating that bias in the point estimate from data mining is probably not serious in our application, but it gets worse as the sample size gets smaller. The mean value is higher for the estimation sample than that of the specification sample. This shows that the specification search leads to better fit and hence a lower value of the optimism. Since the number of correct predictions is higher for specification sample than for the estimation sample, the number of incorrect predictions, conversely, should be lower for the specification sample resulting in lower values of the optimism parameter. Also, the values are consistent with increasing sample size. As the sample size becomes smaller the optimism parameter tends to be higher.

3.5 An Extension

The model presented in this paper could be used to compute regional-average subsidies that would provide the cost of adopting conservation tillage practices. Given, we have four estimates from the four data combinations, the next step is to evaluate which combination is most suitable for this purpose. This section proposes such an extension to the model.

The use of calibration techniques is a well known way to judge how good is a probability estimate. Calibration is a test of whether an issued probability agrees with its relative frequency, *ex post*. The mean probability score or the Brier score is an alternative metric for evaluating probabilistic forecasts which compares the probability of an outcome with the actual outcome. One advantage of Brier score over calibration is that the Brier score can be decomposed into components that index both calibration and resolution, that is the ability of the forecaster to distinguish between events that occur and the events that do not occur.

Let Y be the actual binary outcome of the event. In case of the tillage model, Y takes on the value of 1 of CT is adopted and zero otherwise. Y^* is the probabilistic prediction of the event. Then the quadratic probability score for a single observation or (forecast) is:

$$PS(p, d) = (Y - Y^*)^2 \tag{3}$$

PS ranges between 0 and 1. A score of 0 means perfect prediction, while a score of 1 is bad prediction. This measure is different from the square of the correct predictions.

The mean probability score or Brier score ($\bar{PS}$) is an average of the single prediction version of the the probability score over N occasions, indexed by $i = 1 \ldots N$:

$$\bar{PS}(Y^*, Y) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - Y_i^*)^2 \tag{4}$$

Yates' Covariance Decomposition Calibration does not measure the ability of the forecaster to sort or distinguish between events that actually occur and events that do not occur. The Yates-partition of the Brier score is able to provide information on such sorting. Yates (1982) noted that the mean PS can be factored into its covariance decomposition:

$$\bar{PS}(Y^*, Y) = Bias^2 + Scatter + var(Y) + minvar(Y^*) - 2Cov(Y, Y^*) \quad (5)$$

where, $Var(Y)$ represents the variance of the outcome index, defined as:

$$Var(Y) = \bar{Y}(1 - \bar{Y}) \quad (6)$$

where, $\bar{Y} = 1/N \sum_{i=1}^{N} Y_i$. $Var(Y)$ reflects the factors that are out of the forecaster's control. The remaining terms reflect factors that are under the forecaster's control. In order to obtain the lowest $\bar{PS}$, the forecaster needs to minimize $minvar(Y^*)$, Scatter and $Bias^2$ and maximize $2Cov(Y, Y^*)$.

$$\begin{aligned} Bias &= \bar{Y^*} - \bar{Y} \\ Cov(Y, Y^*) &= Slope * Var(Y) \\ Slope &= \bar{Y}_1^* - \bar{Y}_0^* \end{aligned}$$

where, $\bar{Y}_1^*$ is the conditional mean probability of adopting and $\bar{Y}_0^*$ is the conditional mean probability of not adopting.

$$\begin{aligned} Scatter(Y^*) &= \frac{1}{N}[N_1 Var(Y_1^*) + N_0 Var(Y_0^*)] \\ Var(Y_1^*) &= \frac{1}{N_1} \sum_{i=1}^{N_1} (Y_{i1} - Y_1^*)^2 \\ Var(Y_0^*) &= \frac{1}{N_0} \sum_{i=1}^{N_0} (Y_{i0} - Y_0^*)^2 \end{aligned}$$

Bias quantifies whether the probability predictions are too low or too high. It reflects the overall miscalibration of the forecast. $Bias^2$ reflects the calibration error regardless of the direction of the error.

Scatter is interpreted as an index of general excess variability contained in the forecaster's judgements. The scatter indexes the forecaster's responsiveness to information not related to event's occurrence.

The covariance measures the responsiveness of the forecaster to information related to the event's occurrence. The maximum value of Slope is 1 which occurs when the forecaster always reports $Y_1 = 1$ and the event does

occur and $Y_0 = 0$ and the event does not occur. The covariance term reflects the model's ability to make distinctions between individual occasions in which the event occurs or does not occur.

$Minvar(Y^*)$ is the minimum forecast variance defined as:

$$minvar(Y^*) = Var(Y^*) - Scatter(Y^*) \tag{7}$$

It represents the overall variance in the forecaster's probabilities if there were no scatter about the conditional means $\bar{Y}_1^*$ and $\bar{Y}_0^*$.

In the conservation tillage model, Y^* is the probability of adoption. The actual behavior is given by the variable $Till$

$$Y* = \frac{exp(Estimate)}{1 + exp(Estimate)} \tag{8}$$

Table 6 reports the Brier score for HUC 7080 for each of the four combinations of parameter estimates and data sets. The Brier score for the estimation sample is minimum for specification sample since model uncertainty is least in this case. The specification sample estimation performs the best as it is supposed to, mainly because of the high value of the covariance, reflecting the model's superior ability to make distinctions between individual occasions in which the event occurs or does not occur.

The out-of-sample validation performs marginally better amongst the remaining three estimation types, again mainly because of the covariance term.

Bias is very low for all the estimation types, which indicates an overall good performance of the estimation.

The variance of the actual outcomes Y or the exogenous factors affecting estimations remain more or less constant across the four estimations types.

The scatter terms are highest for the specification and the out of sample estimation. The data set is common in these two cases, which probably explains the general variability in these two models.

The out-of-sample validation estimation performs well when presented under this criteria. Thus the subsidy estimates resulting from these out-of-sample validation would provide reasonable estimates as well as avoid the data-dependent specification search.

4 Conclusions

The objective of this paper is to evaluate the gains of avoiding data-dependent specification search on an estimation sample while estimating a number of conservation tillage adoption models for the Upper Mississippi river basin. We began by splitting randomly the total available data in four sub-samples. We undertook specification search on the specification sub-sample to select the models with the best fit. We then obtained the specification-search-bias-free estimates of model parameters by estimating the models selected on the second, estimation sample. Finally, we used bootstrapping techniques to estimate the measures of excess optimism concerning model fit. We found that the excess optimism is generally small, but varies with the sub-watersheds and has a tendency to be larger for the sub-watersheds with smaller samples.

Because agricultural and ecological data sets are often characterized by large number of observations, the model selection process we followed is viable for these data sets. While we did not find large gains from avoiding the improper specification search in our application, additional research is needed to evaluate the magnitudes of the gains in other applications. An interesting extension of this study would concern evaluating the gains of avoiding data-dependent specification search on the estimation of region-average subsidies needed for adoption of conservation tillage. As the estimates of the conservation tillage adoption model are affected by the specification search, so are the estimates of the subsidies which are the functions of the data and the adoption model parameters.

Table 1: Description of the UMRB watershed by 4 digit HUC

4 Digit HUC	Total cropland points	Total area in million acres	Percentage of total area under cropland	Percentage of cropland area under corn	Percentage of cropland area under soybean	Percentage of total area under conservation till
7010	8954	1.2	18	61	4	2
7020	7797	0.92	69	50	28	12
7030	4113	0.46	10	67	1	2
7040	6495	0.65	33	69	6	14
7050	3847	0.55	11	70	1	4
7060	5930	0.55	42	78	6	32
7070	5141	0.66	14	66	1	5
7080	14965	1.46	67	62	24	45
7090	7167	0.66	56	78	9	22
7100	8375	0.9	64	54	28	43
7110	5883	0.59	44	35	19	14
7120	7661	0.63	55	58	22	18
7130	9745	1.13	72	57	29	26
7140	7776	0.79	44	42	19	13

Table 2: Descriptive Statistics

Notation	Description	Units	Mean	Standard deviation
Y	Conservation tillage(1-yes, 0-no)	Binary number	0.35	0.47
IC	Dummy variable for corn (1-corn,0-not corn)	Binary number	0.59	0.5
IS	Dummy variable for soybean (1-soybean ,0-not corn)	Binary number	0.34	0.47
π_{CVT}	Net returns to conservation tillage	$ per acre	88.88	81.62
SLOPE	Land slope	Percent	2.99	3.2
PM	Soil permeability	Inches per Hour	1.32	1.87
AWC	Available water capacity of soil	Percent	0.21	0.03
PH	Soil acidity (0 to 14) 7-neutral, less than 7 - acidic, greater than 7 - alkaline	Number	6.51	0.50
OM	Plant and animal residue in soil	Percentage	4.35	6.01
EI	Erodibility Index ($EI \geq 8$ are considered highly erodible land)	Number	5.75	9.78
TMAX	Mean of daily maximum temperature during growing season	Fahrenheit	78.5	2.82
TMIN	Mean of daily minimum temperature during growing season	Fahrenheit	55.4	2.95
PRECIP	Mean of daily precipitation during growing season	Inches	0.13	0.01
PRSTD	Standard deviation of precipitation	Inches	0.31	0.03
OFFARM	Proportion of operators working off-farm to the total number of farm operators in the county	Number	0.5	0.06
TENANT	Proportion of harvested cropland operated by tenants to the total county harvested cropland	Number	0.18	0.07
AGE	County average farm operator age	Years	50.71	1.74
MALE	Proportion of male operators to the total number of farm operators in the county	Number	0.97	0.01
CODE	Rural code for counties (0 to 9) 9 completely rural	Code	5.3	2.4

Table 3: Model specification and estimation

HUC	7010	7030	7050	7060	7070	7080
INTERCEPT	-4602.71 (2092.38)	2643.45 (835.1)	1449.5 (596.26)	9845.19 (6421.66)	-1344.09 (1644.82)	3400.85 (1500.53)
CORN ID	15.33 (10.68)	5.2 (3.60)	10.38 (4.15)	21.04 (17.14)	33.64 (15.54)	6.32 (5.72)
SOY ID	14.98 (11.02)	4.2 (3.7)	11.55 (4.42)	17.36 (15.57)	34.89 (16.21)	4.44 (5.73)
SLOPE	-1.98 (1.47)	1.8 (0.6)	1.3 (0.33)	5.39 (3.79)	2.49 (1.08)	1.83 (0.90)
PM	-1.33 (1.14)	-0.8 (0.72)	x	-2.41 (2.60)	x	-0.59 (1.04)
AWC	7.25 (54.70)	-31.9 (38.13)	x	-192.07 (176.45)	x	-94.85 (64.03)
EI	2.11 (1.34)	-0.32 (0.2)	x	-1.55 (1.13)	x	-0.31 (0.28)
OM	-0.01 (0.28)	-0.07 (0.16)	x	0.32 (0.56)	x	0.11 (0.23)
PH	-4.03 (3.97)	3.01 (2.00)	x	5.38 (6.79)	x	0.52 (2.78)
TMAX	-5.39 (2.76)	0.14 (0.6)	x	10.55 (7.11)	x	0.25 (0.94)
TMIN	6.20 (3.48)	2.23 (0.7)	x	-4.68 (3.98)	x	1.20 (1.02)
PRECIP	-12.97 (401.89)	1118.9 (228.6)	1145.44 (230.74)	3134.21 (1963.02)	2204.4 (857.4)	1243.24 (378.59)
TENANT	x	55.3 (100.25)	x	995.44 (683.507)	x	256.19 (193.78)
OFFARM	x	52.9 (105)	x	-1049 (831.25)	x	59.81 (230.74)
AGE	x	-3.6 (3.7)	x	-24.21 (17.51)	x	-1.55 (5.25)
MALE	4740.74 (2145.77)	-2896.5 (827.2)	-1649 (62)	-9539 (6269.92)	1089.8 (1632.9)	-3796.85 (1523.02)
CODE	x	8.6 (2.7)	x	13.90 (11.83)	x	14.44 (5.35)
VPRECIP	-44780.9 (20743)	28914.2 (8446)	14013 (5915.81)	105135 (68027.8)	-14357.2 (16943.3)	35780.70 (14916.50)
VRETURNS	-0.29 (0.21)	0.35 (0.26)	0.27 (0.27)	-0.063 (0.71)	0.8 (0.7)	-0.48 (0.43)
VTENANT	x	297.4 (1019.8)	x	9942.8 (6726.62)	x	2537.93 (1879.31)
VOFFARM	x	447.1 (1040.5)	x	-9626.09 (7813.82)	x	292.73 (2087.97)
VAGE	x	-50.1 (36.8)	-14178 (6066.1)	-292.74 (206.35)	x	-54.01 (51.77)
VMSHARE	45622 (21043.1)	-27495 (8143.8)	x	-90194.5 (59624.6)	15262.9 (17411.1)	-35052.1 (14614.40)
VCODE	x	88.5 (27.4)	x	216.829 (159.48)	x	152.15 (55.84)
Invsigma	14.68 (6.93)	13.7 (2.6)	16.42 (3)	43.80 (28.83)	36.29 (14.38)	17.38 (5.47)

Table 3: Model specification and estimation..continued

HUC	7090	7100	7110	7120	7130	7140
INTERCEPT	7742.55 (7592.82)	8187.26 (2573)	1932.89 (4948.64)	483.212 (417.64)	1825.86 (594.76)	2851.59 (922.25)
CORN ID	86.96 (85.88)	0.72 (6.46)	4.58 (8.46)	13.94 (8.05)	17.21 (7.87)	20.26 (6.96)
SOY ID	102.59 (100.07)	3.21 (6.49)	0.30 (3.58)	14.003 (8.18)	15.04 (7.65)	15.26 (6.16)
SLOPE	8.40 (8.05)	0.40 (0.57)	1.15 (2.10)	3.60 (0.87)	3.85 (1.06)	2.27 (0.71)
PM	x	-0.15 (2.00)	2.52 (4.18)	x	x	x
AWC	x	-76.05 (68.74)	142.96 (241.51)	x	x	x
EI	x	-0.05 (0.18)	-0.04 (0.28)	x	x	x
OM	x	-0.24 (0.56)	1.39 (2.58)	x	x	x
PH	x	-1.70 (1.77)	-1.67 (3.67)	x	x	x
TMAX	x	5.37 (1.32)	2.42 (6.19)	1.46 (0.73)	x	1.87 (0.69)
TMIN	x	-7.13 (1.56)	-7.59 (12.32)	x	x	x
PRECIP	6430.86 (6149.12)	630.07 (196.55)	-226.08 (1060.02)	1305.36 (286.47)	1318.26 (331.28)	2285.15 (593.49)
TENANT	x	569.60 (291.01)	-502.92 (1220.52)	x	x	x
OFFARM	x	347.68 (221.71)	-1229.81 (1173.72)	x	x	x
AGE	x	2.66 (11.35)	-36.86 (64.87)	x	x	x
MALE	-8804.91 (8521.06)	-9023.52 (2762.26)	1266.04 (2530.36)	-791.50 (435.93)	-2048.74 (633.83)	-3320.78 (1004.06)
CODE	x	28.26 (9.75)	-22.71 (18.62)	x	x	x
VPRECIP	45345.2 (56301.8)	81129.00 (25530.00)	6045.21 (40530.3)	3817.34 (4382.54)	23497.9 (7203.35)	36904.2 (10327.3)
VRETURNS	-2.87 (7.34)	-0.18 (0.26)	-9.52 (1.04)	0.13 (0.22)	-0.73 (0.22)	-1.09 (0.29)
VTENANT	x	5006.36 (2575.12)	-9276.06 (16185.7)	x	x	x
VOFFARM	x	1802.43 (2125.39)	-12706.9 (11795.7)	x	x	x
VAGE	x	0.82 (96.73)	-316.16 (599.85)	x	x	x
VMSHARE	-45890.6 (57824.8)	-86034.20 (26886.6)	22212.6 (30870.7)	-3554.98 (4488.29)	-23554.6 (7307.5)	-36694.6 (10410.9)
VCODE	x	245.67 (87.42)	-279.493 (233.44)	x	x	x
Invsigma	125.28 (116.35)	9.53 (2.57)	6.86 (11.18)	34.08 (7.37)	35.65 (9.14)	26.73 (7.10)

Table 4: Goodness-of-fit measures

HUC	7010	7020	7030	7040	7050	7060	7070	7080	7090	7100	7110	7120	7130	7140
Area combinations that best fits the HUC	7010 7020 7030	7010 7020 7030	7030 7040 7080	7030 7040 7080	7040 7050 7080	7060 7080	7060 7070 7080	by itself	7080 7090	by itself	by itself	7080 7120 7130	7080 7130 7140	7080 7110 7130 7140
Model type	D	D	D	D	D	D	D	D	NR	D	NR	D	D	D
N	246	750	77	420	67	406	119	1641	680	856	412	660	1161	580
PCP with specification data	95.1	89.8	88.31	76.85	71.21	62.56	76.47	66.97	68.48	75.23	87.83	66.36	64.43	75.99
PCP with estimation data	95.53	87.15	89.61	75.47	50.74	66.25	84.03	67.64	65.73	74.65	83.49	66.51	63.74	73.45
out-of-sample PCP with specification and estimation data	93.49	87.15	93.5	75.47	68.65	61.33	84.87	61.24	66.17	72.31	85.19	64.7	63.74	73.79
out-of-sample PCP with estimation and validation data	95.95	86.24	85.9	74.52	69.11	61.67	81.6	69.36	66.47	75.38	80.34	66.87	61.53	71.55

PCP: percentage of correct predictions, D: Model specification with difference between the two returns, NR: Model specification with net returns from conventional tillage

Table 5: Bootstrap estimation of the measure of optimism

Sample	Average error	Mean	Std. Dev.	Min	Max
Specification 7080	0.330	0.010	0.011	-0.028	0.05
Estimation 7080	0.323	0.012	0.011	-0.025	0.044
Specification 7100	0.248	0.015	0.013	-0.05	0.07
Estimation 7100	0.25	0.015	0.014	- 0.03	0.06
Specification 7110	0.12	0.019	0.016	-0.05	0.07
Estimation 7110	0.16	0.025	0.018	- 0.03	0.08

Table 6: Yates Decomposition of the Brier Score

Estimation Types	Brier Score	Bias Square	Variance of Till	Covariance	Scatter	Minimum variance of prediction
Specification	0.1975	0.000	0.2483	0.0513	0.0413	0.0106
Validation	0.2089	0.000	0.2479	0.0388	0.0326	0.0061
Out-of-sample	0.2207	0.00003	0.2479	0.039	0.0447	0.0061
Out-of-sample validation	0.2011	0.0002	0.2473	0.0435	0.0329	0.0076

References

[1] C. Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society*, Series A(158):419–466, 1995.

[2] M. Creel and J. Loomis. Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in california. *American Journal of Agricultural Economics*, 72, 1990.

[3] B. Efron. *The jackknife, the Bootstrap, and other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics, 1982.

[4] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

[5] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap.* London: Chappman and Hall, 1993.

[6] Herriges J. Egan, K., C. Kling, and J. Downing. Water quality as a function of physical measures. *Woking Paper*, 2005.

[7] J. Faraway. Data splitting strategies for reducing the effect of model selection on inference. *Computing Science and Statistics*, 30, 1998.

[8] G. Gong. Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of American Statistical Association*, 81(393):108–113, 1986.

[9] Kling C. Kurkalova, L. and J. Zhao. Green subsidies in agriculture: Estimating the adoption costs of conservation tillage from observed behavior. *Working Paper 01-WP-286, CARD, ISU*, 2003.

[10] E. Leamer. Let's take the con out of econometrics. *American Economic Review*, 73(1):31–43, 1983.

[11] Roberts M.C. Norwood, B. and J. L. Lusk. Ranking crop yield models using out-of-sample likelihood functions. *American Journal of Agricultural Economics*, 86, 2004.

[12] R. Picard and R. Cook. Cross-validation of regression models. *Journal of Americal Statistical Association*, 79, 1984.

[13] B. M. Potscher. Effects of model selection on inference. *Econometric Theory*, 7(2), 1991.

[14] M. R. Veall. Bootstrapping the process of model selection: An econometric example. *Journal of Applied Econometrics*, 7(1), 1992.

[15] M. R. Veall and K.F. Zimmermann. Pseudo-r^2 measures for some common limited dependent variable models. *Journal of Economic Surveys*, 10:241–59, 1996.

[16] F.A. Yates. External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, 30, 1982.

[17] P. Zhang. Inference after variable selection in linear regression model. *Biometrika*, 79, 1992.