Interpreting Results of Demand Estimation from Machine Learning Models

Gareth P. Green (corresponding author)
Chair, Department of Economics
Albers School of Business and Economics
Seattle University
Seattle, WA 98122
Email: greeng@seattleu.edu

Timothy J. Richards
Morrison School of Agribusiness and Resource Management
Arizona State University
Mesa, AZ 85212

Selected Paper prepared for presentation for the 2016 Agricultural & Applied Economics Association, Boston, MA, July 31-August 2

Introduction

There is developing interest in the application of Machine Learning Models (MLM) to estimation problems in economics (Varian 2014 and Belloni, Chernozhukov and Hansen 2014). MLM may be particularly well suited to applications in retail, health care, energy, finance or for web based businesses where large amounts of data are available to help make better decisions and better understand consumer behavior. Varian suggests three reasons economists may want to adopt new MLM tools. First is the size of available data sets. While historically economists may have felt starved for data to model behavior, now there are a wide array of massive data sets of market and social behavior. Second, these new data sets have many potential predictors where domain knowledge may not be helpful in distinguishing which available data are most relevant. For example, when considering purchase of a specific product, domain knowledge will help a general group of substitutes; however, for each individual the data may now indicate which specific product serves as a substitute for the specific product under consideration. Third, larger data sets allow for modeling more complex relationships than the standard linear model, which is what MLM are able to capture.

Though engineers, computer scientists and data scientists have long been familiar with MLMs (Donoho 2015), most economists are unfamiliar with them. There is a distinct difference in the application of MLM and standard econometric models. Data analysts that use MLMs are generally more interested in building prediction models where there is little interest in interpretation and statistical validity. Rather, data scientists divide data into training and testing sets, estimate different models with the training sets, then validate the models on the testing data. The model with the best prediction results is selected as the preferred model. Because of this approach, economists that are familiar with MLMs have sometimes voiced skepticism because

they are not based on domain expertise and do not provide interpretable coefficients or measures of statistical validity. However, MLMs have been shown to have better out-of-sample predictive capabilities than many standard econometric tools (Bajari et.al. 2015), so likely have a place in the economist's tool set for predicting and forecasting.

We develop MLM and econometric prediction models and compare prediction results from MLMs to several standard econometric models of demand estimation on data that have large numbers of variables and observations. These types of data sets are becoming more prevalent and also are *an area where standard econometric models do not perform well*. We take a new approach to simulate coefficients for interpretation of results, but recognize that MLM do not have standard normal asymptotics so the interpretations are open to criticism. However, we find that even without standard normal asymptotics we can still develop models that predict better than standard econometric models and provide coefficients for interpretation. We use an approach suggested by Varian (2014), first use MLM prediction models to estimate a base case scenario. Second, find the difference between the predicted base case and actual behavior under a treatment. The difference indicates the estimated impact of the treatment on behavior.

We use IRI Marketing Research scanner panel data on cold cereal from a set of grocery stores (Bronnenberg, Kruger, Mela 2008). When we include product and store fixed effects, the models have thousands of explanatory variables and over a million observations. Determining which explanatory variables to use based on domain expertise in this case is near futile and parameters from standard regression models are poorly estimated. As a result, variable selection is an important problem. We use several MLM's that focus on reducing dimensionality, including stepwise regression, forward stagewise regression, LASSO and support vector machines (Belloni, Chernozhukov, Hansen 2014). We also employ two common regression tree

models, bagging and random forests. We use the conditional logit and panel data regression models as standard econometric models to compare with the MLMs. All analysis is done in the open source statistical package R using specialized packages. R and the specialized packages are freely available, well-documented and have been tested in many applications (CRAN 2015).

Models and Methodology

We begin with a general demand modeling framework because the estimation model for each methodology is slightly different. Quantity demanded for a set of *j* products is modeled as:

(1) $Q = \sum_{jt} q_{jt} = f(p_{jt}, x_j, a_{jt}, y_i, \varepsilon_{ijt})$,

where p_j, x_j and a_j are the price, product and promotional attributes of good *j*; y_i are demographic attributes of consumer *i*; and ε_{ij} is a random error. We assume quantity, price and promotional attributes vary over time and product and consumer attributes are constant; which are constraints imposed by the data set. The model is very general, which allows us to use a variety of estimation specifications.

We begin by estimating the demand model using nine different methods: linear regression, logit, logit with boosting, forward stepwise regression, forward stagewise regression, LASSO regression, random forest, support vector machine, and L2 boosting regression. The nine models we employ have been examined extensively in the computer science literature (Ahmed et al. 2010, Wen et al. 2012). A brief review of each MLM used here, including asymtotics, has also been given in Bajari et al. (2015). A brief description of each type of method is given in Table 1.

We will compare the performance of each model by dividing the data set into three parts: a training set, a testing set and a validation set. The training set is used to develop and tune the model to get the best performance. The models are then applied to the testing set to insure the

model is not "over fitting" the training set. That is, it is possible to train a model that achieves excellent prediction results on the training data set; however, performs poorly on out-of-sample prediction. Once the models are finalized they are all run on the validation data to judge the performance of each model. We use standard error (SE) as our primary measure of model performance.

Next we will apply all nine models to estimate the impact of promotional programs on product sales. There are two critical issues for judging the impact of a promotional program when you cannot run a controlled experiment: 1) sample selection bias – the products and type of promotion are likely to be different across products in a way dependent specific on the product; and 2) determining the "causal" impact of the promotional program—how much did sales change from what they would have been. The first issue arises from it simply being expensive and unlikely to run controlled and randomized experiments within and across products. The second arises from not know "what sales would have been" if the promotion were not run. We use prediction models to address these two issues.

Our approach will follow that suggested by Varian (2014). First we will estimate a model on data with no promotion activity which will be used to forecast the counterfactual product sales—the "sales that would have been" without the promotion. The model also captures sample selection bias characteristics related to products, markets and seasons. The better the prediction model is, the better the baseline will be. Next we will use a second set of data that includes both non-promotional and promotional activities to predict what sales would have been without the promotional program. Then we will examine the difference between the predicted and actual product sales to estimate the impact of the promotional programs, both individual observations and for brands as a total. We could also do this for specific stores or time periods. Finally, we

calculate the per unit impact of the promotion on sales to get an inferential measure of the promotion similar to a coefficient. Distributions of the coefficients could also be generated using Monte Carlo simulation, though that is not included in this study.

Application and Results

We apply the models and methodology to the IRI Marketing Research scanner panel data on cold cereal from a set of grocery stores, see Bronnenberg, Kruger, and Mela (2008) for a detailed description. We select this application because it includes the large dimensionality issues inherent in data sets that are becoming readily available and is an area where traditional econometric techniques struggle. When we include product and store fixed effects, the models have thousands of explanatory variables and over a million observations. For example, in the cold cereal data set there are 1,299 different stores, 86 brands, 31 different flavors and 22 grain types. To make the data set more manageable (and not require high-power computing), we do not examine store fixed effects and only look at the most common brands, flavors and grain types (the top fifty-percent). While this obscures the relative higher performance of the MLM compared to the traditional models, one would not likely use a traditional model with that many variables.

The models are estimated using packages in R, an open source statistical program. The package used to estimate each model is listed in Table 1. Most of the models require some level of tuning to increase the model performance in terms of prediction precision, yet the tuning has to be weighed against the problem of over-fitting the model to the training data. The models is then applied to a validation data set (that has not been used to develop or test the model) after arriving at a final model for each method to measure its final performance. We show the performance in terms of SE for each model on each of the data set in Table 2.

The SE is larger for all models on the testing and validation sets than on the training data, as expected. The Forward Stagewise and LASSO have the smallest increase in SE in the validation set relative to the training set, and similarly have the lowest SE across the models. Though the and L2 Boosting has the largest increase in SE across all models, it still has one of the lowest SE. The Random Forest model does not perform the worst in the validation, which is surprising as we have seen it to do very well in other applications. The linear regression does surprisingly well relative to the other models, which has not been the case in other studies (Bajari et al. 2015). This could be due to our reduction of the dimensionality to reduce solution time. Bottom line, there is still a lot to learn about these models and how they perform in different settings.

A summary of the promotional lift is given in Table 3. The first column indicates the mean of the difference between the actual data and the counterfactual prediction. The models that performed the best in the validation test tend to show the highest mean difference, or largest impact, of the promotional programs. That would indicate that the other models estimate a higher counterfactual case relative to the actual data. That is, the linear, Stepwise and Random Forest models appear to produce upward biased forecasts, so would underestimate the impact of the promotional program on product sales. The mean difference by brand is larger than across all observations because some brands do very little promotion while others do a significant amount. The same pattern of mean difference is evident in the brand specific mean difference as across all observations, as would be expected.

Moving forward there are many questions to examine. We have yet to complete the Monte Carlo simulations to derive coefficients and their distributions. We can allow for more explicit substitution effects into the model by including competitors' prices as variables in the

quantity equation. We can examine specific brands to see how they are impacted by promotions for other products and vice versa. Changes in revenues will be calculated to find the benefit of the promotional programs. Finally, some of the modeling techniques used here are adept at incorporating variables like packaging or promotional placement that can have a non-linear impact on sales. Moving forward we will delve deeper into these questions to illustrate the strengths and weaknesses of using machine learning models in economic applications.

Conclusions

Though the approach we are suggesting is controversial among some econometricians, we feel it makes sense to use lots of data based on actual behavior rather than rely purely on domain expertise that in these cases may not be able to discern an efficient set of variables and statistical tests that depend on randomness. Further, these methods make it possible to use unstructured data based on text, which can easily influence consumer behavior. The application possibilities of MLM will continue to grow as more transactions and behavior are tracked with scanners and sensors, so economists should become more knowledgeable of their benefits and limitations. We use a familiar application to demonstrate how MLMs compare and contrast to standard economic models of demand estimation as a basis for discussion of begin building this knowledge.

References

Ahmed, N. K., A. F. Atiya, N. El Gayar, and H. El-Shishiny. 2010. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews*, 29(5–6):594–621.

Bajari, P, D. Nekipelov, S. Ryan, M. Yang. 2015. Demand estimation with machine learning and model combination. *National Bureau of Economic Research,* Accessed at www.nber.org/papers/w20955. Accessed October 8, 2015.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. High dimensional methods and inference on structural treatment effects. *Journal of Economic Perspectives 28*(2): 29-50.

CRAN. 2015. The Comprehensive R Archive Network. Accessed at https://cran.r-project.org/ on November 17, 2015.

Donoho, D. 2015. 50 years of data science. Accessed at http://www.r-bloggers.com/50-years-of-data-science-by-david-donoho/ on December 14, 2015.

Bronnenberg, B., M. Kruger, C. Mela. 2008. Database paper: The IRI marketing data set. *Marketing Science*, 27(4) 745-748.

Varian, H. 2014. Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2): 3-28.

Wen, J., S. Li, Y. Hu and C. Huang. 2012. Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology 54*(1):41-59.

Table 1: Machine Learning Model Descriptions

Linear Regression	Estimate models linear in components, but not necessarily in variables, R package {stats}
Logit	Estimate dichotomous outcome variables, R package {stats}
Logit with Boosting	Calculates a gradient booting object to find optimal number of trees for logistic regression, R packages {stats}{dismo}
Forward Stepwise Regression	Automated choice of predicative variables based on correlation with response variable, R package {lars}
Forward Stagewise Regression	Automated choice of predicative variables based on correlation with the residual, R package {lars}
LASSO Regression	High dimension tuning intensive regression used to reduce dimensionality, similar to ridge regression, R package {flare}
L2 boosting Regression	Efficient boosting algorithm with l2-loss function and design matrix columns, R package {l2boost}
Support Vector Machine	A discriminative classifier used in regression or classification defined by separating hyperplanes, R package {e1071}
Random Forest	Classification and regression based on generating random forests of trees, R package {randomForest}

Table 2: Root Mean Square Error

Model	Training	Testing	Validation
Linear Regression	0.012642	0.019807	0.018683
Logit	0.013959	0.018861	0.018964
Logit with Boosting	0.013198	0.017634	0.019962
Forward Stepwise Regression	*wip*	0.018312	0.018248
Forward Stagewise Regression	*wip*	0.015846	0.013635
LASSO Regression	0.009118	0.012730	0.011411
L2 boosting Regression	0.006246	0.014184	0.013287
Support Vector Machine	0.010795	0.019309	0.019074
Random Forest	0.012455	0.019350	0.020822

wip – work in progress

Table 3: Promotional Lift

Model	Mean diff all observations	SD of difference	Mean diff across brands	SD of diff across brands
Linear Regression	0.081587	0.926257	2.628815	9.13414
Logit	*wip*	*wip*	*wip*	*wip*
Logit with Boosting	*wip*	*wip*	*wip*	*wip*
Forward Stepwise Regression	0.081865	0.926616	2.637772	8.95879
Forward Stagewise Regression	0.094417	0.926517	3.042202	9.04252
SQRT LASSO Regression	0.114747	0.965749	3.697245	13.12068
L2 boosting Regression	0.112915	0.960293	3.638219	12.36502
Support Vector Machine	0.131867	1.059856	4.248865	10.65540
Random Forest	0.092867	0.906211	2.992256	8.54124

wip – work in progress