American Economic Review: Papers & Proceedings 2015, 105(5): 481–485
http://dx.doi.org/10.1257/aer.p20151021

Machine Learning Methods for Demand Estimation[†]

By Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang*

Over the past decade, there has been a high level of interest in modeling consumer behavior in the fields of computer science and statistics. These applications are motivated in part by the availability of large datasets, and are commonly used by firms in the retail, health care, and Internet industries to improve business decisions. In this paper, we compare these methods to standard econometric models that are used by practitioners to study demand.[1] We are motivated by the problem of finding practical tools that would be of use to applied econometricians in estimating demand with large numbers of observations and covariates, such as in a scanner panel dataset.

Many economists are unfamiliar with these methods, so we briefly sketch several commonly-used techniques from the machine learning literature.[2] We consider eight different models that can be used for estimating demand: linear regression, the conditional logit, and six machine learning methods, all of which differ from standard approaches by combining an element of model selection into the estimation procedure. Several of these models can be seen as variants on regularization schemes, which reduce the number of covariates in a regression which receive nonzero coefficients, such as stepwise regression, forward stagewise regression, LASSO, and support vector machines. We also consider two models based on regression trees, which are flexible methods for approximating arbitrary functions: bagging and random forests. While these models may be unfamiliar to many economists, they are surprisingly simple and are based on underlying methods that will be quite familiar. Also, all of the methods that we use are supported in statistical packages. We perform our computations in the open source software package R. Therefore, application of these methods will not require writing complex code from scratch. However, applied econometricians may have to familiarize themselves with alternative software.

We apply our method to a canonical demand estimation problem. We use data from IRI Marketing Research (Bronnenberg, Kruger, and Mela 2008) via an academic license at the University of Chicago. It contains scanner panel data from grocery stores within one grocery store chain for six years. We used sales data on salty snacks, which is one of the categories provided in the IRI data. The number of observations is 1,510,563, which includes 3,149 unique products.

If we allow for product and store level fixed effects, our model effectively has many thousands of explanatory variables. Therefore, variable selection will be an important problem. If we included all of these variables in a standard regression model, the parameters would be poorly estimated. Also, many of the regressors will be multi-collinear which will make the models predict poorly out of sample.

In our results, we find that the six models we use from the statistics and computer science literature predict demand out of sample in standard metrics much more accurately than a panel data or logistic model. We do not claim that these models dominate all methods proposed in the voluminous demand estimation literature. Rather, we claim that as compared to common methods an applied econometrician might use in

* Bajari: Department of Economics, University of Washington, 331 Savery Hall, Seattle, WA 98195, and NBER (e-mail: bajari@uw.edu); Nekipelov: Department of Economics, University of Virginia, 254 Monroe Hall, Charlottesville, VA 22904 (e-mail: dn4w@virginia.edu); Ryan: Department of Economics, University of Texas at Austin, 2225 Speedway Stop C3100, BRB 3.134D, Austin, TX 78712, and NBER (e-mail: sryan@utexas.edu); Yang: Department of Economics, University of Washington, 331 Savery Hall, Seattle, WA 98195 (e-mail: yangmiao@uw.edu).

[†] Go to http://dx.doi.org/10.1257/aer.p20151021 to visit the article page for additional materials and author disclosure statement(s).

[1] Hastie, Tibshirani, and Friedman (2009) is a comprehensive reference.

[2] See, for example, Belloni, Chernozhukov, and Hansen (2014) and Varian (2014).

off the shelf statistical software, these methods are considerably more accurate. Also, the methods that we propose are all available in the well documented, open software package R as well as commercially-available software.

Finally, we propose using an idea dating back at least to Bates and Granger (1969). We treat each of these eight independent predictions as regressors and form a combined model by regressing the dependent variable on to the prediction of each component model. We use a three-way cross validation to avoid overfitting the models in practice. We split the sample into three disjoint sets; we use the first set to fit all eight models, we use the second set to fit our regression on the eight independent model predictions, and we use the third set of the data to test the fit out of sample. We find that this combination procedure can lead to substantial improvements in fit with little additional work. And, as we detail in Bajari et al. (2014), the combined model exhibits standard asymptotic behavior, even though the component models may not, which simplifies the construction of standard errors.

I. Summary of Machine Learning Methods

We explore using machine learning techniques to predict demand. We briefly discuss each method in turn before applying them to estimation of demand using a scanner panel dataset.

A typical specification for demand of product *j* in group *h* in market *m* at time *t* would be

$$(1)\quad Y_{jhmt} = f(\mathbf{X}, \mathbf{D}, \mathbf{p})'\beta + \zeta_{hm} + \eta_{mt} + \epsilon_{jmt},$$

where *f* generates arbitrary interactions between the observables ($\mathbf{X}$), demographics ($\mathbf{D}$), and prices ($\mathbf{p}$). Such a model may have thousands of right-hand-side variables; an extreme example from Rajaraman and Ullman (2011) notes that Google estimates the demand for a given web page by using a model of the network structure of literally billions of other web pages on the right-hand side. Dummy variables on nests are captured by ζ_{hm}. Seasonality is captured by the term η_{mt}, which varies by time (say, quarters) across markets.

In ordinary least squares (OLS), the parameters of equation (1) are typically estimated using the closed-form formula $\beta = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$. This formula requires an inversion of $(\mathbf{X}'\mathbf{X})$, imposing a rank and order condition on the matrix $\mathbf{X}$. We highlight this because in many settings, the number of right-hand-side variables can easily exceed the number of observations. Even in the simplest univariate model, one can saturate the right-hand side by using a series of basis functions of $\mathbf{X}$. This restriction requires the econometrician to make choices about which variables to include in the regression. We will return to this below, as some of the machine learning methods we discuss below allow the econometrician to skirt the order condition by combining model selection and estimation simultaneously.

A large literature on differentiated products has focused on logit-type models, where the idiosyncratic error term is assumed to be distributed as a Type I Extreme Value. Under that restriction, market shares are given by

$$(2)\qquad s_{jhmt} = \frac{\exp(\theta'\mathbf{X}_{jhmt})}{\sum_{k\in J}\exp(\theta'\mathbf{X}_{khmt})}.$$

Quantities are then computed by multiplying through by market size.[3]

Stepwise regression starts with the intercept as the base model on a set of demeaned covariates. The algorithm then searches over the set of covariates, selects the one with the highest correlation with the residual, and adds that variable to the next model. The method then estimates OLS using that subset of covariates, then repeats the search for the covariate with the next highest correlation. The procedure produces a series of nested models and runs until no covariates have a sufficiently high correlation with the error term.

Forward stagewise regression is a variant on the stepwise regression. Whereas all of the coefficients can change at each step in the stepwise regression, forward stagewise regression only updates one coefficient at each step. The method finds the variable with the highest correlation with the error term and adds that covariance to the coefficient. This continues until none of the

[3] Berry, Levinsohn, and Pakes (1995) extend this model to deal with unobserved heterogeneity and vertical characteristics that are observed to both the firm and consumers.

covariates have any correlation with the error term.

These methods build up the model over time in addition to estimating a fit. One advantage of this approach is that the methods can recover the true data-generating process when the number of covariates is larger than the number of observations and the true model is sparse, e.g., the number of coefficients with true nonzero values is less than the number of observations.

Support vector machines (SVM) are a penalized method of regression, using the following:

$$\min_{\beta} \sum_{i=1}^{n} V(y_i - \mathbf{X}_i'\beta) + \frac{\lambda}{2}|\beta|, \tag{3}$$

where the loss function is

$$V_\epsilon(r) = \begin{cases} 0 & \text{if } |r| < \epsilon, \\ |r| - \epsilon & \text{otherwise.} \end{cases} \tag{4}$$

The tuning parameter, ϵ, controls which errors are included in the regression. Errors of sufficiently small size are treated as zeros. Typically only a partial set of the covariates are assigned a nonzero value in SVM regression.

LASSO is another penalized regression method. The regression is given by

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda(t - \sum_{j=1}^{p} |\beta_j|), \tag{5}$$

where t is the tuning parameter governing how strictly additional regressors are penalized. LASSO typically results in a number of covariates being given zero weights.

Regression trees approximate functions by partitioning the characteristic space into a series of hypercubes and reporting the average value of the function in each of those partitions. Regression trees generalize fixed effects to allow them to depend on values of $\mathbf{X}$. In the limit as the hypercubes grow infinitesimally small, the tree reports the average value $Y = f(\mathbf{X} = \mathbf{x})$, which is a perfect reconstruction of the underlying function f. In practice, the tree is expanded until the reduction in squared prediction error falls under some threshold. Often, the tree is grown until a specific number of splits are achieved.

The literature has proposed several variations on the regression tree estimator. One is bagging (Breiman 1996), which uses resampling and model combination to obtain a predictor. The idea is to sample the data with replacement B times, train a regression tree on each resampled set of data, and then predict the outcome at each x through a simple average of the prediction under each of the B trees.

A second approach, which we have found to work exceptionally well in practice, are random forests, as in Breiman (2001). Random forests expand on the idea of using collections of predictors to make predictions by introducing randomness into the set of variables which are considered at node level for splitting. Before each split, only $m \leq p$ of the explanatory variables are included in the split search. Repeating this across B trees results in a forest of random trees. The regression predictor for the true function is then

$$\hat{f}_{rf}^{B}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}). \tag{6}$$

Trees of sufficient size can be unbiased but exhibit high variance, and therefore may benefit from averaging.

II. Empirical Application

This section compares econometric models with machine learning ones using a typical demand estimation scenario—grocery store sales. We find that the machine learning models in general produce better out-of-sample fits than linear models without loss of in-sample goodness of fit. If we combine all the models linearly with nonnegative weights, the resulting combination of models produces better out-of-sample fit than any model in the combination.

The data we use is provided by IRI Marketing Research via an academic license at the University of Chicago. It contains scanner panel data from grocery stores within one grocery store chain for six years. We used sales data on salty snacks, which is one of the categories in the IRI data. A unit of observation is product j, uniquely defined by a UPC (Universal

Table 1—Model Comparison: Prediction Error

	Validation		Out-of-sample		Percent weight
	RMSE	SE	RMSE	SE	
Linear	1.169	0.022	1.193	0.020	6.62
Stepwise	0.983	0.012	1.004	0.011	12.13
Forward stagewise	0.988	0.013	1.003	0.012	0.00
LASSO	1.178	0.017	1.222	0.012	0.00
Random forest	0.943	0.017	0.965	0.015	65.56
SVM	1.046	0.024	1.068	0.018	15.69
Bagging	1.355	0.030	1.321	0.025	0.00
Logit	1.190	0.020	1.234	0.018	0.00
Combined	0.924		0.946		100.00

Product Code), in store m at week t. The number of observations are 1,510,563, which includes 3,149 unique products. Let q_{jmt} be the number of bags of salty snack j sold in store m at week t. If $q_{jmt} = 0$, we do not know if it is due to no sale or out-of-stock and the observation is not filled in. The price p_{jmt} is defined as the quantity weighted average of prices for product j in store m at week t. Therefore if $q_{jmt} = 0$, the weight is also 0. In addition to price and quantity, the data contains attributes of the products (such as brand, volume, flavor, cut type, cooking method, package size, fat and salt levels) and promotional variables (promotion, display, and feature).

The response variable is log of quantity sold per week. The covariates are log of price, product attributes variables, promotional variables, store fixed effects, and week fixed effects. We provide the same covariate matrix to all of the models except for the logit model, where all the fixed effects are excluded.[4]

In order to estimate and compare models, we split our data into three sets: training, validation, and holdout. We estimate the model using the training set, and then use the validation set to assign weights to each model when building the combined model. This approach mitigates overfitting in the training model; for example, the linear model tends to get a good in-sample fit but a bad out-of-sample fit, and granting these models a large in-sample weight would produce poor predictions in the holdout sample. Finally, we combine each model to predict fit in a holdout sample. Twenty-five percent of the data is used as the holdout sample, 15 percent is used as the validate set, and the remaining 60 percent is used as the training set.

Table 1 reports the root mean squared prediction error (RMSE) across the validation and out-of-sample datasets, along with the estimated weights of each model in the combined model. Judged by out-of-sample prediction error, the best two models are random forest and support vector machine. The combined model, where we regress the actual value of the response variable on a constrained linear model of the predictions from eight models, outperforms all the eight models, which follows the optimal combination of forecasts in Bates and Granger (1969). Random forest receives the largest weight in the combined model (65.6 percent), and the stepwise and forward stagewise models receive the majority of the rest. It is interesting to observe the combined model does not simply choose the submodel with the best RMSE; there are important covariances among the models which generate better fit in combination than any one given submodel.

III. Conclusion

In this paper, we review and apply several popular methods from the machine learning literature to the problem of demand estimation. Machine learning models bridge the gap between parametric models with user-selected covariates and completely non-parametric approaches. We demonstrate that these methods can produce superior predictive accuracy as compared to a standard linear regression or logit model. We also show that a linear combination of the underlying models can improve fit even further with very little additional work. While

[4] For brevity, we have minimized the description of the dataset and details of implementation of the machine learning methods; Bajari et al. (2014) provides more details about the construction of the dataset.

these methods are not yet commonly used in economics, we think that practitioners will find value in the flexibility, ease-of-use, and scalability of these methods to a wide variety of applied settings.

One concern has been the relative paucity of econometric theory for machine learning models. In related work (Bajari et al. 2014), we provide asymptotic theory results for rates of convergence of the underlying machine learning models. We show that while several of the machine learning models have non-standard asymptotics, with slower-than-parametric rates of convergence, the model formed by combining estimates retains standard asymptotic properties. This simplifies the construction of standard errors for both parameters and predictions, making the methods surveyed here even more accessible for the applied practitioner.

REFERENCES

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. 2014. "Demand Estimation with Machine Learning and Model Combination." Unpublished.

Bates, John M., and Clive W. J. Granger. 1969. "The Combination of Forecasts." *Operational Research Quarterly* 20 (4): 451–68.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28 (2): 29–50.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Equilibrium." *Econometrica* 63 (4): 841–90.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–40.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Bronnenberg, Bart J., Michael W. Kruger, and Carl F. Mela. 2008. "Database paper–The IRI Marketing Data Set." *Marketing Science* 27 (4): 745–48.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Rajaraman, Anand, and Jeffrey D. Ullman. 2011. "Mining of Massive Datasets." Lecture Notes for Stanford University CS345A Data Mining. http://infolab.standard.edu/~ullman/mining/2009/index.html.

Varian, Hal R. 2014. "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives* 28 (2): 3–28.