An empirical study of predicting car type choice in Sweden using cross-validation and feature-selection

Shiva Habibi Marcus Sundberg
Anders Karlström

KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden

Keywords: hold-out sample, out of sample prediction, feature selection, cross validation, model selection, car type choice, discrete choice modeling, clean vehicles.

The composition of the car fleet concerning age, fuel consumption and fuel types has a great impact on environment. Recently, several different policies have been implemented to affect this composition. Therefore, building models that predict the future composition of car fleet more reliably is very important. These models are used to predict and evaluate the impacts of these policies on the fleet to provide policy makers with decision supports. In this paper we analyze the prediction problem and focus on building a multinomial logit model (MNL) to predict accurately, the market shares of new cars in the Swedish car fleet in the short-term future. Also, we investigate whether or not different prediction questions lead to different 'best' models.

We use feature (variable) selection and cross-validation algorithms to improve predictive performance (Stone, 1974; Geisser, 1975; Efron and Morris, 1973; Guyon and Elisseeff, 2003). Feature selection is an automatic way of selecting variables to be included in the model such that the criterion which is called loss function is optimized. This method is very useful in car type choice application since there exists a large number of car attributes to select among as well as their correlations and interactions. Selecting variables based only on priori knowledge is not likely to be accurate (see e.g. the Train, 1979). Cross-validation (CV) is an accuracy estimation method in which data is split, once or several times, part of the data (the training sample) is used for estimation (training), and the remaining part (the validation sample) is used for validating the estimated results. CV selects the models which give the smallest error over validation sample. A single data split is called simple validation or hold-out validation, and averaging over several splits is called cross-validation. CV improves prediction by avoiding over-fitting. Over-fitting occurs when a model is too fitted to the available data that looses its generality to be applied on another independent data.

In the car type choice modeling area, employing a multinomial logit (MNL) model to predict the influence of transport policies starts with the work of Lave and Train, 1979. However, in this area, as other fields of choice modeling, there are only few studies focusing on evaluating the predictive accuracy of models. In an effort to predict future demand for clean cars,

Brownstone et al., 1994, develop a forecasting system based on microsimulation. Attributes of future vehicles are exogenous to this system. They apply a bootstrapping method to measure the effect of the forecast error. Mohammadian and Miller, 2002 and Hensher and Ton, 2000 compare the predictive potential of nested logit (NL) with artificial neural network (ANN) in car-type choice application. Cross-validation method is used in these studies in neural network framework whereas standard estimation is used for Logit, hence one may question the results when comparing NN-Logit in that the results seen in those studies might be the effect of different methodologies (CV vs. standard estimation). Interestingly, there is no study that applies CV method in logit models in car type choice application and to the best of our knowledge, no study uses feature selection to choose variables to be included in the car type choice models. In this paper, we employ feature selection and cross-validation to select the MNL model of car type choice which provides the highest predictive performance.

We use Swedish car fleet data for the years 2006-2008 and our objective is to build a model to predict year 2008 market share based on data on years 2006 and 2007. Since we have access to time-series data We build two types of models, using 2007 cross-sectional data and also pooled data 2007-2006 as training samples. We suggest the hypothesis that using pooled data 2006-2007 will result in more robust models to predict year 2008 data. On the demand side, we have the Swedish car register containing data on all registered cars together with some characteristics such as brand, model, weight and horse power. Moreover, we have data on car alternatives available in the market for 2006-2008, denoted by supply. These data are , down to a specific version of a model. For 2006, 2007 and 2008 there are respectively, 2320, 2679 and 2981 cars alternatives, corresponding to 45 different makes. We need to map each observation from demand to each alternative in supply. Since supply is more detailed than the registry data, several alternatives can correspond to a given observation. Therefore, we need to aggregate these alternative. We use the method introduced in Ben-Akiva and Lerman, 1985 to aggregate alternatives. Considering the changes in the supply of each year, using the supply of a given year both for estimation and validation is not likely to give us accurate predicted results since the model might be over-fitted to the supply of that year. Therefore, we use hold-out validation where the validation sample is the data of consecutive year to assess the predictive performance of a given model. We introduce four different loss functions associating with different prediction questions: log-likelihood, root mean square error for brands market share, root mean square error for ethanol (E85)/brand market share, and total number of ethanol cars. The predicted results of log-likelihood as a loss function are compared with those of the remaining ones.

The results show that different loss functions result in different 'best' models. In other words it shows that different prediction questions lead to different 'best' models. However, the predicted results reject our hypothesis about obtaining higher predictive performance using the pooled data. It can be said intuitively that this result shows the sensitivity of models to supply in that the supply of 2008 is more similar to 2007 one than to 2006 which suggests the probable over-fitting of the models to the supply of two successive years (i.e. years 2007 & 2008). An example predicted results for market share of E85 cars are presented in following table. Comparing the results of different loss function indicates the predicted results of 'best' models obtained from the associated loss function on 2008 data, outperform the 'best' model resulted from the log-likelihood as a loss-function. This can be explained by the fact that LL assigns the

same weight to all alternatives and consequently gives the overall prediction whereas in prediction we are usually interested in a sub-section of the data *e.g.* E85 cars. Therefore; the objective of selecting the models with good overall predictive performance, as in log-likelihood, is not the same as the objective of selecting the models aiming at predicting more focused and specific question such as total share of E85 cars.

		Estimation on 2007; Validation on 2008			**Estimation on 2006-2007; Validation on 2008**		
Loss function:		LL	RMSE E85 share	Total E85	LL	RMSE E85 share	Total E85
Brand	**Actual share**		**Predicted share**			**Predicted share**	
CADILLAC	0.10	0.12	0.01	_	0.02	0.06	_
CHEVROLET	0.02	0.01	0.09	_	0	0.09	_
CITROEN	0.17	1.21	1.22	_	0.34	0.72	_
FORD	3.04	2.55	3.13	_	0.88	3.03	_
PEUGEOT	1.16	2.16	1.44	_	1.22	1.74	_
RENAULT	0.76	1.87	0.52	_	0.26	0.77	_
SAAB	3.27	4	2.9	_	0.73	2.27	_
SEAT	0.25	0.22	0.12	_	0.12	0.22	_
SKODA	1.33	1.73	1.37	_	1.33	1.01	_
VOLKSWAGEN	1.95	1.39	1.6	_	1.47	2.19	_
VOLVO	7.83	2.1	7.8	_	3.1	7.84	_
Total	**19.89**	17.36	20.2	**19.89**	**9.47**	19.94	**19.9**
RMSE		**1.86**	**0.37**		**1.76**	**0.41**	

Table 1: Predicted results for market share of E85 cars

References

Ben-Akiva, M. and Lerman, S. (1985). *Discrete choice analysis: theory and application to travel demand*, Vol. 9, MIT press.

Brownstone, D., Bunch, D. S. and Golob, T. F. (1994). A Demand Forecasting System for Clean-Fuel Vehicles, *Transportation* (221): 15.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its CompetitorsAn Empirical Bayes Approach, *Journal of the American Statistical Association* **68**(341): 117–130.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications, *Journal of the American Statistical Associatioin* **70**(350): 320–328.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection, *The Journal of Machine Learning Research* **3**: 1157–1182.

Hensher, D. A. and Ton, T. T. (2000). A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice, *Transportation Research Part E: Logistics and Transportation Review* **36**(3): 155–172.

Lave, C. A. and Train, K. (1979). A disaggregate model of auto-type choice, *Transportation Research Part A: General* **13**(1): 1–9.

Mohammadian, A. and Miller, E. J. (2002). Nested Logit Models and Artificial Neural Networks for Predicting Household Automobile Choices: Comparison of Performance, *Transportation Research Record: Journal of the Transportation Research Board* **1807**(1): 92–100.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2): pp. 111–147.

Train, K. E. (1979). A comparison of the predictive ability of mode choice models with various levels of complexity, *Transportation Research Part A: General* **13**(1): 11–16.